

DOCUMENT RESUME

ED 346 731

FL 020 440

TITLE Foreign Language Testing. Supplement 1981-1987.
Specialised Bibliography 6.
INSTITUTION Centre for Information on Language Teaching and
Research, London (England).
REPORT NO ISBN-0-948003-82-0
PUB DATE 88
NOTE 146p.
PUB TYPE Reference Materials - Bibliographies (131)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Annotated Bibliographies; *Applied Linguistics;
*Communicative Competence (Languages); Elementary
Secondary Education; Foreign Countries; *Language
Tests; Scholarly Journals; *Second Language
Instruction; Testing; *Testing Problems; Trend
Analysis

ABSTRACT

A specialized bibliography on foreign language testing compiled by the British Centre for Information on Language Teaching and Research (CILT) is presented covering the period 1981 to 1987. This supplement consists of 200 abstracts from the journals "Language Teaching and Linguistics: Abstracts" and "Language Teaching." It also contains an annotated listing of the following: 4 survey articles; 11 sources of information on testing and examinations; nearly 60 books, including conference reports, surveys, monographs, and research reports; 55 selected titles of articles that do not appear in the abstract section; and 10 examples of published tests and additional collections of test materials held by the CILT Library. A detailed subject index and name index are also included. The literature reflects substantial interest in testing problems involved in specifying communicative proficiency and in reshaping the public examination system at 16+. (LB)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

FL 020 440

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

John
Hawkins

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

Specialised Bibliography 6

FOREIGN LANGUAGE TESTING

SUPPLEMENT 1981-1987

Compiled and edited at CILT

CENTRE FOR INFORMATION ON LANGUAGE TEACHING AND RESEARCH

Copyright ©1988 Centre for Information on Language Teaching
and Research
ISBN 0 948003 82 0

Printed in Great Britain by Staples Printers Rochester Ltd.

Published by Centre for Information on Language Teaching and
Research, Regent's College, Inner Circle, Regent's Park, London
NW1 4NS

All rights reserved. No part of the publication may be reproduced,
stored in a retrieval system, or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording, or other-
wise, without the prior permission of the Copyright owner.

Contents

	Page
1. Introduction	1
2. Abstracts 1-200	2
3. Survey articles S1-S4	109
4. Bibliography B1-B115	110
4.1. Sources of information on testing and examinations B1-B11	110
4.2. Books (Monographs, Conferences, Surveys, Research Reports) B12-B71	111
4.3. Indexes and periodicals B72-B77	125
4.4. Periodical articles B78-B115	126
5. Tests	131
5.1 Examples of published tests T1-T10	131
5.2 Collections of test materials held by the CILT Library	133
6. Name index	134
7. Contents index of Abstracts	135

Introduction

INTRODUCTION TO THE SUPPLEMENT TO THE SECOND EDITION

The first and second editions of this bibliography published by CILT in 1977 and 1981 respectively covered work on testing recorded by the *Abstracts* in Volumes 1-13, 1968-1980.

The supplement consists of abstracts from Volume 14 (1981) of the *Language Teaching and Linguistics: Abstracts* and from Volumes 15 to 20 (1982-1987) of *Language Teaching** together with an annotated listing of survey articles, books and other items considered to be appropriate, published mainly during the same period. During this time interest in testing has focussed mainly on the problems involved in specifying communicative proficiency and in the reshaping of the public examination system at 16+. A detailed subject index and name index completes the volume.

Books and journals may usually be consulted in academic libraries. When not available, a loan/photocopy service is provided to UK and overseas library users by the British Library Document Supply Service (Boston Spa, Wetherby, Yorkshire LS23 7BQ) through the International Photocopy Service.

CILT wishes to thank the Syndicates of Cambridge University Press for again allowing abstracts of articles to be reprinted from *Language Teaching*.

* *Language Teaching* (formerly *Language Teaching and Linguistics: Abstracts* 1975-1981, and *Language Teaching Abstracts* 1968-1974) is a quarterly journal, compiled by the British Council and the Centre for Information on Language Teaching and Research, and published by Cambridge University Press. It provides summaries in English of selected articles appearing in about 400 British and foreign journals covering all aspects of language teaching and applied linguistics.

Abstracts

1. ALDERSON, J Charles

The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly* (Washington, DC), 13, 2 (1979) [publ. 1982], 219-27.

The cloze test has received considerable attention in recent years from testers and teachers of English as a foreign language, and is becoming more widely used in language tests, both in the classroom and in standardised tests. However, most of the research has been carried out with native speakers of English and the results do not produce clear-cut evidence that the cloze test is a valid test of reading comprehension. The article reports on a series of experiments carried out on the cloze procedure where the variables of text difficulty, scoring procedure and deletion frequency were systematically varied and that variation examined for its effect on the relationship of the cloze test to measures of proficiency in English as a foreign language. Previous assumptions about what the cloze procedure tests are questioned and it is suggested that cloze tests are not suitable tests of higher-order language skills, but can provide a measure of lower-order core proficiency. Testers and teachers should not assume that the procedure will produce automatically valid tests of proficiency in English as a foreign language.

2. ALDERSON, J Charles, and Alan Waters

A course in testing and evaluation for ESP, or 'How bad were my tests?' *Lancaster Practical Papers in English Language Education* (Lancaster), 5 (1982), 39-61.

There is a need in ESP for evaluation and recognition of its performance, but a lack of acceptable evaluation instruments. In ESP projects and teacher training generally, testing is relegated to a minor role, while materials and method training loom large. Since experience shows all teaching to be influenced by terminal tests, good or bad, the development of good ESP tests could lead to better ESP teaching and materials. Such tests should be of use in the classroom and appeal to the teacher, and not be remote and forbidding as current 'testing' elements in teacher training courses make them appear. Teachers need not become testing experts, merely intelligent and sensitive evaluators.

To produce such, a seminar was given at Lancaster University which was interactive, problem-based and classroom-oriented, covering eight activity topics dealing with testing issues and the writing, scoring and interpretation of tests, culminating in an evaluation of the seminar itself. The response was overwhelmingly positive. The idea that ESP teaching is healthy but its testing weak is challenged, since teaching and testing are reciprocal and poverty in one implies poverty in the other.

3. ALDERSON, J Charles

Testing, the teacher and the student. *Cahiers de l'APLIUT* (Paris), 39, 3 (1984), 7-21.

Testing is regarded with distaste by many teachers and apprehension by many students, not always for very good reasons. Conversely, testers do their cause no good by claiming more for their tests than is justified and separating themselves by arcane statistics from the teaching community. The various reasons for this divorce are examined in turn.

Testers should be more concerned with what they are measuring than the nature of measurement; they should test what teachers want to teach and what learners want to learn. Since tests will always be subjective, if not in their marking then in their construction, they should go hand in hand with teaching to sample worthwhile behaviour, e.g. speaking, representing worthwhile educational goals. By consciously testing for the washback effect on teaching the relationship between the teaching and testing community will be a positive and mutually beneficial one. For teachers who have to compose tests, 10 general precepts and 19 specific instructions are given.

4. ANTIER, Maurice

M'entendez-vous bien? - peut-on mesurer la compréhension auditive? [Do you understand? Can listening comprehension be measured] *Langues Modernes* (Paris), 78, 4 (1985), 313-23. Complete issue entitled 'L'évaluation'.

Language teachers have become increasingly interested in listening comprehension during the last 15 years. From a review of the literature during that period (with particular reference to English) it is concluded that: listening comprehension and comprehension of written texts call for different skills and abilities, though there appears to be some correlation between listening comprehension and rapid reading; specialised language should be avoided in testing listening comprehension; and some so-called

comprehension tests are really memory tests. Points of special difficulty to francophone learners are identified.

The tester should use authentic language spoken in a natural way, check different aspects of listening comprehension separately (stress, intonation, segmental phonemes, grammatical structures, vocabulary) as well as checking overall comprehension. A variant of the cloze test with key words omitted is suggested.

5. ARNAUD, Pierre J L

La place de la dictée dans une batterie d'entrée. [The place of dictation in an entrance test battery.] *Langues Modernes* (Paris), 78, 4 (1984), 281-91.

In the present era of integrative tests and under the impulse of Spolsky and Oller, dictation has enjoyed a revival and stands, with cloze, as a valid and reliable test of underlying language proficiency. As practised by Spolsky against a background of noise, it is a reduced redundancy test quite different from the classic dictation, in which sound conditions are optimised. High correlations between cloze, dictation and TOEFL are reported, but it is clearly not a communicative test.

The author describes the entrance test he administered at the University of Lyon and correlates dictation scores with those on the other three elements - vocabulary, grammar and essay. While the dictation correlated at 0.71 with the total score on the other tests, none correlated highly enough with any of the others to justify its sole use and the predictive validity of the whole test can be assessed only after the students' sessional results are known.

6. ATZE, Charlotte

Zur Bewertung und Zensurierung von mündlichen Leistungen. [The evaluation and marking of oral performance.] *Fremdsprachenunterricht* (Berlin, GDR), 25, 10 (1981), 479-86.

Suggestions are made for a concrete system of criteria for the objective evaluation and marking of spoken language. To have a firm basis for such a marking system, the teacher must first of all decide exactly what he is expecting of the pupils and the pupils must also be aware of what is expected of them. The teacher must also distinguish between the types of language production: whether it is monologue or dialogue and whether it is guided or free. Marking criteria must vary accordingly. Detailed guidelines are provided for marking, taking into account various factors, in the form of descriptions of criteria for evaluation and evaluation tables.

7. BABCOCK, Arthur

Proficiency testing in French at the University of Southern California. *French Review* (Baltimore Md), 58, 3 (1985), 403-10.

A proficiency test for students of French is described and its rationale explained. An examination was needed showing level of achievement in French as distinct from coursework and it was decided to have a reading comprehension test accompanied by a conversation test.

The conversation test was a new and encouraging departure based on magazine pictures, which candidates had to describe to three separate and successive assessors. They were given a grading on a ten-point scale for global conversational proficiency and the interrator reliability coefficient proved very high (0.815 to 0.941). Good results were achieved with the reading test and both together show improved statistics on each successive administration. It not only provides an accurate measure of French proficiency but also has a beneficial washback effect on teaching and learning, encouraging concentration on the most worthwhile skills.

8. BACHMAN, Lyle F, and Adrian S Palmer

The construct validation of the FSI oral interview. *Language Learning* (Ann Arbor, Mich), 31, 1 (1981), 67-86.

The validation of oral language proficiency tests has long been problematical. The most common approaches to this problem have involved concurrent validation procedures for relating 'indirect' measures to so-called 'direct' measures. The most frequently used criterion in such studies has been the Foreign Service Institute (FSI) oral interview. In validating the FSI interview itself, however, criterion-referenced procedures are not possible since an adequate operationally defined criterion for comparison is not likely to be found. Therefore construct validation procedures are necessary.

In this study a multitrait-multimethod matrix comprising six measures representing combinations of two traits (speaking and reading) and three methods (interview, translation, self-rating) was used to examine the construct validity of the FSI oral interview. The six measures were administered individually to 75 native Mandarin Chinese speakers of English. Data were analysed using the Campbell-Fiske criteria for convergent and discriminant validity and using confirmatory factor analysis.

The results indicate both convergent and discriminant validity for the FSI interview. Of the numerous factor models tested against the data, the model which provides the best fit comprises a general factor, two trait factors, and three method factors. A more

parsimonious model, with two correlated trait factors and three method factors, fits the data nearly equally well and provides a more interpretable factor structure. These results provide strong evidence against a model of unitary language competence and support a model of partially divisible language competence.

9. BACHMAN, Lyle F

The trait structure of cloze test scores. *TESOL Quarterly* (Washington, DC), 16, 1 (1982), 61-70.

Although there is considerable evidence supporting the predictive validity of cloze tests, recent research into the construct validity of cloze tests has produced differing results. Chihara et al. (1977) concluded that cloze tests are sensitive to discourse constraints across sentences, while Alderson (1979) concluded that cloze tests measure only lower-order skills. Anderson (1980) has concluded that cloze tests measure sensitivity to both cohesive relationships and sentence-level syntax. Factor analytic studies (Weaver & Kingston, 1963; Ohnmacht et al., 1970) have identified several factors in cloze and other language tests and suggest that cloze deletions should be based on the linguistic and coherence structures of language.

In the present study, the trait structure of a cloze test was examined using confirmatory factor analysis. A cloze passage with rationally selected deletions of syntactic and cohesive items was constructed and given to two groups of non-native English-speaking students entering the University of Illinois. A trait structure with three specific traits and one general trait provided the best explanation of the data. The results suggest that a modified cloze passage, using rational deletions, is capable of measuring both syntactic and discourse level relationships in a text, and that this advantage may outweigh considerations of reduced redundancy which underlie random deletion procedures.

10. BACHMAN, Lyle F, and Adrian S Palmer

The construct validation of some components of communicative proficiency. *TESOL Quarterly* (Washington, DC), 16, 4 (1982), 449-65.

The notion of communicative competence has received wide attention in the past few years, and numerous attempts have been made to define it. Canale and Swain (1980) have reviewed these attempts and have developed a framework which defines several hypothesised components of communicative competence and makes the implicit claim that tests of components of communicative competence measure different abilities. In this study the construct

validity of some tests of components of communicative competence and of a hypothesised model is examined. Three distinct traits - linguistic competence, pragmatic competence and sociolinguistic competence - were posited as components of communicative competence.

A multitrait-multimethod design was used, in which each of the three hypothesised traits was tested using four methods: an oral interview, a writing sample, a multiple-choice test and a self-rating. The subjects were 116 adult non-native speakers of English from various language and language-learning backgrounds. Confirmatory factor analysis was used to examine the plausibility of several causal models, involving from one to three trait factors. The results indicate that the model which best fits the data includes a general and two specific trait factors - grammatical/pragmatic competence and sociolinguistic competence. The relative importance of the trait and method factors in the various tests used is also indicated.

11. BACHMAN, Lyle F

Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly* (Washington, DC), 19, 3 (1985), 535-55.

Although research on the cloze test has offered differing evidence regarding what language abilities it measures, there is a general consensus among researchers that not all the deletions in a given cloze passage measure exactly the same abilities. An important issue for test developers, therefore, is the extent to which it is possible to design cloze tests that measure specific abilities.

Two cloze tests were prepared from the same text. In one, different types of deletions were made according to the range of context required for closure, while in the other a fixed-ratio deletion procedure was followed. These tests were administered to 910 university and pre-university students, including both native and non-native speakers of English, with approximately half assigned at random to take the fixed-ratio test and the other half taking the rationally deleted test.

While both tests were equally reliable and had equal criterion validity, the fixed-ratio test was significantly more difficult. Analyses of responses to different types of deletions suggest that the difficulty of cloze items is a function of the range of syntactic and discourse context required for closure. The study also provides practical and empirically supported criteria for making rational deletions and suggests that cloze tests can be designed to measure a range of abilities.

12. BENSOUSSAN, Marsha

Testing the test of advanced EFL reading comprehension: to what extent does the difficulty of a multiple-choice comprehension test reflect the difficulty of the text? *System* (Oxford), 10, 3 (1982), 285-90.

This paper examines the faulty assumption that the difficulty of a multiple-choice comprehension test depends mainly on the difficulty of the text. Instead, the basic assumption stated in this paper is that the difficulty of a multiple-choice test depends on two components, a text and questions, with results varying according to the level of English proficiency of the students tested. Two multiple-choice comprehension tests on the advanced level were administered: one based on a relatively easy text, the other on a more difficult one. The difficulty of the texts was measured by teacher evaluation and by discourse analysis. For each text, nine multiple-choice questions are examined. Test 1 was administered to 178 candidates, and Test 2 to 132. The two tests show no significant differences in scores, despite the fact that one of the texts is clearly easier. The results indicate that teachers/test-constructors cannot assume that a text of a particular level of difficulty will automatically yield questions reflecting the level of the text.

13. BENSOUSSAN, Marsha

Dictionaries and tests of EFL reading comprehension. *ELT Journal* (London), 37, 4 (1983), 341-5.

The use of dictionaries in EFL (English as a Foreign Language) tests is a controversial issue. Test constructors and administrators wish to refrain from introducing any extraneous materials into the test situation, whereas teachers would like to incorporate dictionaries because they are seen to be an integral part of the reading process. The aim of this investigation is to focus on some of the underlying assumptions by which dictionaries may or may not be permitted during examinations. Towards this end, test questions were especially written, bearing in mind that students would be using dictionaries. The rationale for these questions is explained; they ask about the function or effect of words and sentences in context. It was concluded that the dictionary does not really 'give' the student the meaning of a word. Readers have to negotiate that for themselves according to the context.

14. BENSOUSSAN, Marsha, and others

The effect of dictionary usage on EFL test performance compared with student and teacher attitudes and expectations. *Reading in a Foreign Language* (Birmingham), 2, 2 (1984), 262-76.

This paper reports the results of three studies, designed to discover whether the use or non-use of dictionaries had an effect on students' performance on EFL reading tests. The general conclusion is that, when given the choice, students preferred to use bilingual dictionaries, but that the use of dictionaries had no significant effect on students' test scores. The results of a questionnaire administered to both students and teachers to ascertain dictionary use and preferences is also reported. The better students preferred not to use a dictionary in a test.

15. BENSOUSSAN, Marsha, and Rachel Ramraz
Testing EFL reading comprehension using a multiple-choice rational cloze. *Modern Language Journal* (Madison, Wis), **68**, 3 (1984), 230-9.

A multiple-choice rational cloze dubbed a 'fill-in test' is shown to be as reliable in scoring as the ordinary multiple-choice test while being easier to construct, more adaptable, economical and more easily targeted on a particular aspect of language to be tested.

The problems of multiple-choice and cloze testing are reviewed and the linguistic basis of the fill-in test discussed. The designers left 20-30 spaces in a 300-word text; the omissions were not random but chosen because of their key function in a rational argument, which the candidate had to reconstruct. Four different multiple-choice and fill-in tests were used on populations of over 50 students and the results compared and correlated. Results correlated highly but the fill-in test proved more reliable, partly because of the larger number of items tested. Because of these results the fill-in test is now being used to test about 13,000 students annually for entrance to the universities of Haifa and Tel Aviv.

16. BERNHARDT, Elizabeth B
Testing foreign language reading comprehension: the immediate recall protocol. *Unterrichtspraxis* (Philadelphia, PA), **16**, 1 (1983), 27-33.

Traditional methods of testing reading comprehension are problematic. True/false and multiple-choice tests are often not passage-dependent (i.e. random answers score as well as considered ones). In any case, it is difficult to know what prior knowledge the reader already has. Designing comprehension questions is a subjective matter. Perhaps the only true test of reading a foreign language is a translation test, since it staves off the problems of test construction. Teachers working within a communicative syllabus, however, defeat their own purpose if they test with translation or cloze, since these demand explicit grammatical competence.

The immediate-recall protocol procedure tests students' abilities to understand written foreign language texts without the help of outside materials. Students read a passage silently as often as they like, surrender the text and write down in English everything they remember from the text. [Sample passage and three protocols from college level students of German; sample scoring instrument]. A recall protocol reflects process rather than product, and shows how the subjects' cognitive processes work. Without testing for grammar it shows where a lack of grammar is interfering with the communication which should be going on between students and text. It in no way influences the students' understanding of the text, and it stresses the importance of understanding the material. This encourages students to monitor themselves. It is easy to construct and administer.

17. BIALYSTOK, Ellen

On the relationship between knowing and using linguistic forms. *Applied Linguistics* (Oxford), 3, 3 (1982), 181-206.

Traditionally language proficiency is assessed in quantitative terms as a number of items known by the learner; improvement is regarded as a process of increasing these items. However, it is shown in two experiments described in detail that language proficiency can and should be assessed in qualitative terms; more specifically, as a relation between this knowledge of forms and the contexts of use that mobilise or neutralise it. Variability in performance among language learners who know exactly the same formal items derives from their differential control over two factors: (1) analysed representations of knowledge, and (2) automatic access to information; (roughly, conscious knowledge of how the language works, and retention in the mind of a sufficient number of items). The experiments show that where native speakers, advanced and intermediate foreign learners are tested on structures they all know in common, the higher scores can be accounted for by an ability to use them flexibly, subject to an increasing range of situational demands. Modality (oral or written) is excluded as a factor, leaving only analysed and automatic control as the decisive elements. Hence instruction must consider the specific goals of the learners and anticipate contexts of use for their knowledge.

18. BLUE, Walter, and others

An alternative approach to oral proficiency testing: the practical exam. *French Review* (Baltimore, Md), 58, 2 (1984), 185-90.

The testing or oral proficiency by one-to-one interview is excellent but time-consuming and the article describes a test in French which allows a group of 30 students to be tested in less than half the time needed for interviews, each of them speaking French for 25 minutes with 5 different examiners. The candidates are given a situation framework and instructions on how they have to act in it. Thus they have to talk their way out of problems that might beset them in France, taking five minutes for each of five tasks. The five examiners' grades are averaged to give a grade for each student. Completion of the task successfully achieves the minimum grade. The exam is popular, practical and a convincing demonstration of ability to cope in France.

19. BOGAARDS, Paul, and Tom Duijkers

Nadere onderzoekingen naar de taal attitude schaal. [Further investigations into the language attitude scale (TAS).] *Toegepaste Taalwetenschap in Artikelen* (Amsterdam), 17, 3 (1983), 132-50.

Contradictory results obtained with TAS by Spoelders (1978) and Bogaards (1982) induced them to make further investigations into this language attitude scale. Two problems were addressed in particular. (i) What is the predictive value of the TAS in respect to school results, and (ii) Does the TAS measure one or more aspects of language attitude? The TAS was administered to 561 students in 25 classes (brugklas: 12-13 years old; 3 havo: 14-15 years old (comprehensive school); 3 vwo: 14-15 years old (grammar school); 5 vwo: 16-17 years old (grammar school). Analyses of the results show that the predictive value of the TAS with respect to school results is considerably higher than that mentioned by Spoelders (0.30 vs 0.18). The TAS measures two aspects, namely 'fun' and 'usefulness'; a test that has been reduced to 14 items seems to measure these aspects at least as accurately as the complete TAS.

20. BOLZ, Norbert, and others

Die Gestaltung remedialen Grammatikunterrichts auf der Oberstufe mit Hilfe von Diagnosetests. [The role of diagnostic tests in the remedial teaching of grammar in upper secondary education.] *Die neueren Sprachen* (Frankfurt am Main), 81, 4 (1982), 377-85.

Teachers of English often have difficulty in determining their pupils' mastery of English grammar. In this article, experiments with a Diagnostic English Grammar Test are reported. This test, which covers 16 areas, was given to eleventh form pupils. A computer output provided the teacher with exact information about each pupil's level of knowledge. Based on this information, a therapy

programme was carried out in an attempt to correct the pupils' weaknesses.

21. BÖTTCHER, Karl-Heinz

Neue Formen der mündlichen Abschlußprüfung im Fach Englisch - entwickelt und erprobt an der Realschule Lauenburg (Schleswig-Holstein). [A new type of oral exam for the school-leaving examination in English - developed and tried out at the 'Realschule' in Lauenburg (Schleswig-Holstein).] *Englisch* (Berlin, FRG), 17, 3 (1982), 87-91.

The oral examination represents an important hurdle for many candidates for the Realschulabschluß (secondary-school leaving qualification). In order to give the weaker pupil a chance, however, the oral should test his communicative ability in English and not his ability to talk in abstractions about the language, as in the written examination. Moreover, the pupil should be prepared for the oral by being given the opportunity to speak English in class, and the form of the examination should allow him to draw on what he has learnt and 'stored'.

A model oral which meets these criteria, was developed and tried out in Lauenburg under the auspices of the ISPT (Schleswig-Holstein schools research institute). Here, pupils not only engaged in question and answer with the teacher/examiner, but also conversed among themselves. An example of the model is set out here and details of the procedure, stages and content are given, as well as suggestions for alternative approaches. Also provided are marking schemes based on (A) linguistic competence, (B) communicative competence. It is the examiner's job to decide how the marking should be weighted.

22. BROWN, David

Conversational cloze tests and conversational ability. *ELT Journal* (Oxford), 37, 2 (1983), 158-61.

Can written cloze tests help us to measure a language learner's oral ability? Arthur Hughes, in *ELT Journal* 35, 2, suggested that 'conversational' cloze tests (which are cloze tests constructed from the transcripts of conversations) can do this. This article considers a subsequent piece of research designed to examine further the relationship between performance on conversational cloze tests and oral ability. Thirty students of English language each took part in two interviews, which were recorded. They also completed two conversational cloze tests. The interviews and the cloze tests were then scored. The ranks of the students' interview scores were then

compared with the students' ranks on the cloze test scores. The results showed a high correlation, which suggests that conversational cloze tests may be a good measure of oral ability.

23. BROWN, James Dean

Relative merits of four methods for scoring cloze tests. *Modern Language Journal* (Madison, Wis), 64, 3 (1980), 311-17.

Four methods of scoring cloze tests are compared: exact-answer (EX), acceptable-answer (AC), Clozentropy (CLZNT) and multiple-choice (MC). It was hypothesised that the four scoring methods would not differ in reliability, validity, mean item facility and discrimination, or in usability. A cloze passage was prepared in an open-ended and a multiple-choice format, each of which was administered to about 50 students of similar language background. Results showed that there were differences in reliability, mean item facility and discrimination, and usability, but not in validity [table ranks the four scoring methods for each criterion]. More studies are needed of what cloze tests are really testing. The best overall scoring method is the AC method, which seems fairer to the students than the EX method.

24. BUCKBY, Michael

Graded objectives and tests for modern languages: an evaluation. *British Journal of Language Teaching*, 19, 1 (1981), 13-14 and 33.

A summary of the findings of a research project on graded objectives and tests carried out at the Language Teaching Centre at the University of York in 1979-80. Two experimental groups were formed, one in Leeds and one in York, with corresponding control groups in which pupils continued with French in the normal way. Most pupils were aged 13+. All pupils, parents and teachers first completed an attitude questionnaire, and later in the year took an aptitude test. At the beginning of the year the attitudes and aptitudes of experimental and control pupils were almost identical: the full ability range was represented. Parental support was at the same level for all pupils.

At the end of the year, pupils, parents and teachers completed the attitude questionnaires again, after the experimental pupils had taken a graded test, learned the results and, where appropriate, received a certificate. These pupils showed very significantly more positive attitudes to learning French than the control pupils, at all points of the aptitude scale. Results in Leeds and York were similar, but gains were more marked in York schools, where graded tests had been used for three years. More experimental pupils expressed a

wish to continue French than did control pupils. The experimental group's parents likewise tended to become more positive over the year, whereas the control group's remained unchanged or became less positive. This suggests that the attitudes of children and parents influence each other. Both groups of teachers seemed well disposed towards graded syllabuses and tests, and believed they helped both the higher and lower ability ranges.

25. BYRAM, Michael

Where is the 16+ leading us? *British Journal of Language Teaching*, 20, 3 (1982), 145-8.

It is pointless to try to motivate pupils to learn a foreign language by reference to their future needs, which scarcely exist in the real world, especially because this means turning away from the heart of language teaching as a part of general education. The documents relating to the 16+ examination system contain no reference to pupils acquiring an understanding of language, only to 'gaining insight into language learning'. The diffuseness of the formulation is a reflection of two factors in the malaise of language teaching: the question of what is testable or examinable (only language behaviour being testable), and the question of methods for achieving aims other than communicative skills, such as 'language awareness' and 'bilingual vision' (seeing the foreign culture from the inside). Much could be done to develop the teaching of cultural studies. 'Communicative competence' is not the destination, only a station.

26. CANALE, Michael

Communication: how to evaluate it. *Bulletin of the CAAL* (Montreal), 3, 2 (1981), 77-94.

Five criteria in judging the usefulness and quality of any good test are validity, reliability, practicality, acceptability and feedback potential. An adequate test of communication must satisfy two main conditions: (1) it must be based on sound descriptions of communication and proficiency in communicating, or 'communicative competence'; (2) these descriptions must be reflected not only in the test content (what is to be measured) but also in test method (how content is to be measured), i.e. test formats, administration and scoring procedures.

Two main ways in which evaluation of communication can be made more practical are (a) to reduce the time and cost involved in test administration and scoring, and (b) to try to develop tests that address core rather than peripheral aspects of communication and

from which the results would be maximally generalisable. Concrete suggestions along these lines include group testing, which seems suitable for assessing oral communication skills, and 'holistic' (impressionistic) scoring with multiple judges.

27. CARROLL, Brendan J

Specifications for an English language testing service. *ELT Documents* (London), 111 (1981), 66-110.

The Council's existing testing system has become unsatisfactory both because (a) the size of the placement problem has grown enormously - the number of applicants for training in Britain has grown and the range of courses of study has expanded; and (b) because of changing approaches to test development language teaching and testing methods have shifted their emphasis from atomistic language features to broader features of linguistic communication, postponing consideration of language realisations until the communicative needs of all users have been clearly determined. The way to improve the testing service is to diversify the test instruments to meet the wide range of test situations. The questions which need answering about any student are whether he is likely to be able to meet the communicative demands of his course, or how much tuition he would need to bring him up to the required level.

A specification is being compiled to build up profiles of the communicative needs of students on study programmes in Britain [the parameters are described]. Specifications have been prepared for six areas: business studies, agricultural science, medicine, social survival (academic), civil engineering, and laboratory technician. [Appendix gives specifications.] They cover purpose of study, settings for English (surprisingly uniform), interactions, instrumentality (receptive/productive skills, spoken written media), dialect (usually contemporary English and local varieties), target level (size and complexity of text), events and activities (what participants have to do), attitudinal tone index (recognition and production), language skills (54 categories), and microfunctions.

The implications for test design are that language skill requirements cut right across disciplinary boundaries: important factors are (1) general, i.e. common communicative requirements, (2) academic study, i.e. ability to use skills necessary for handling academic discourse, (3) personal relationships, and (4) special additional requirements of odd-man-out programmes. A framework for measurement intelligible to non-specialists is suggested. Operational requirements of overseas representations are discussed:

these include availability of tests, quick results, a rapid screening device, cost and security.

Recommendations are that a two-level testing pattern should be adopted with a quick screening test followed by a modular test pattern covering skills appropriate to six major areas. A 'shadow' test pattern will be tried out.

28. CARROLL, Brendan J

The implications for testing of the communicative approach to language teaching. *Studies in Language Learning* (Urbana-Champaign), 4 (1982), 45-55.

Proposals are made for improving the status of communicative testing based on defining the purposes and needs of the testee and describing the behaviour required of him. First, a distinction is made between communication/use and language/usage; the focus of testing should be the former not the latter, though results on both will tend to correlate highly. Statistics and correlations should be treated with caution since much of any correlation is an artefact of the testing instrument used. Indeed, until the goals and the terminal behaviour of a course are clear, the statistical results will be equivocal. Testing is not purely for measurement but for its washback effect on courses, motivation, credibility, teaching and learning strategies. No off-the-peg test can therefore be wholly adequate for each course or group of testees. Overemphasis on statistics often obscures specification faults and leads to the rejection of valid test elements. Simulation and role play should be encouraged to give psychological authenticity to tasks, and results should be based on behaviour band descriptions expressed numerically. Further research into all these desiderata is required.

29. CHAVEZ-OLLER, Mary Anne, and others

When are cloze items sensitive to constraints across sentences? *Language Learning* (Ann Arbor, Mich), 35, 2 (1985), 181-206.

This article reviews the continuing controversy about whether cloze items in general are sensitive to constraints beyond 5-10 words on either side of them. It is concluded that some are sensitive to constraints that reach even beyond 50 words on either side of a blank. The research shows that this sort of context sensitivity is inaccessible to rank beginners, but becomes increasingly accessible to more proficient language users. In addition to reviewing several successful replications of these findings together with several failed attempts, this article gives an item-by-item analysis of the experimental data from Chihara et al. (1977). It is demonstrated

that about 10 per cent of the items in each of the two texts examined are highly sensitive to constraints ranging across sentence boundaries. The basis for this sensitivity is discussed and the sensitive items are contrasted with non-sensitive ones.

30. CHIHARA, Teturo, and others

Are cloze items sensitive to constraints across sentences? *Language Learning* (Ann Arbor, Mich), 27, 1 (1977) [publ. 1979], 63-73.

To determine whether or not cloze items are sensitive to constraints across sentences, two passages of prose were selected and two types of tests were constructed over each passage. In the sequential type, items were inserted in the normal prose by a standard word-deletion procedure. In the scrambled type, the sentences of each passage of prose were systematically shuffled. Each cloze item appeared once in the sequential condition and once in the scrambled. The design was counterbalanced for passage and order of testing. Forty-one native speakers of English and 201 Japanese adults studying English as a foreign language were tested. Each subject took a cloze test over one of the passages in the sequential condition and the other in the scrambled condition. A four-way analysis-of-variance sustained the hypothesis that cloze items embedded in normal prose are sensitive to discourse constraints ranging beyond the immediate limits of a single sentence. There was a significant interaction between level of proficiency and condition indicating that the contrast between the two conditions increases with an increase in proficiency.

31. CLARK, John L D

Language testing: past and current status - directions for the future. *Modern Language Journal* (Madison, Wis), 67, 4 (1983), 431-43.

After a review of the recent history of language testing in terms of the 'three ages' proposed by Spolsky - pre-scientific, psychometric-structuralist and integrative-sociolinguistic - attention is drawn to current trends within the last of these paradigms. The first is 'direct proficiency' testing as exemplified in the Foreign Service Institute interviews and the associated rating scales. These measure performance in language tasks approximating to real life, but usually only within a polite register. The second is the 'hybrid tests' designed by Omaggio which combine grammar and context, structure and situation, and enable the teacher to test the reduced corpus taught up to any point in a course within the real-life situations where it might be mobilised. The third is the use of computers for diagnostic assessment as well as teaching, which

would take away much of the time-consuming work of preparing and marking tests and allow for flexible response to a cumulative profile of candidate performance. Increased research into curriculum-free direct proficiency testing is called for, along with development of computer-assisted diagnostic testing and the orientation of teachers towards explicit proficiency goals.

32. COHEN, Andrew D, and Elite Olshtain
Developing a measure of sociocultural competence: the case of apology. *Language Learning* (Ann Arbor, Mich), 31, 1 (1981), 113-34.

For a number of years, there has been interest in measuring sociocultural competence (Ervin-Tripp, 1972; Hymes, 1974; Canale & Swain, 1980). This study focuses on one important aspect of such competence: the ability to use the appropriate sociocultural rules of speaking, by reacting in a culturally acceptable way in context and by choosing stylistically appropriate forms for that context. We chose to look at productive performance in sociocultural aspects of speaking, focusing on the speech act of 'apology'. The research question that prompted this study was, 'Can a rating scale be developed for assessing sociocultural competence?.'

The subjects were 32 native Hebrew speakers, 20 of whom served as informants for apologies in English L2 and 12 as informants in Hebrew L1, and 12 Americans who served as informants in English L1. These subjects were asked to role-play their responses in eight situations in which an apology was expected. The findings show that it is possible to identify culturally and stylistically inappropriate L2 utterances in apology situations. The authors feel, however, that the results so far provide at best a crude measure of sociocultural competence and that further work with this speech act and with others is called for.

33. COHEN, Andrew D
On taking language tests: what the students report. *Language Testing* (London), 1, 1 (1984), 70-81.

This paper discusses methods for obtaining verbal report data on second-language test-taking strategies; reports on types of findings obtained in a number of studies conducted by university students on how learners take reading tests, in particular cloze and multiple-choice tests; and finally considers the implications of the findings for prospective test takers and test constructors. The main conclusion is that a closer fit should be obtained between how test constructors intend their tests to be taken and how respondents actually take them.

34. COLLETT, M J

Examples of applications of computers to modern-language study. 1. The step-wise development of programmes in reading, grammar and vocabulary. *System* (Oxford), 8, 3 (1980), 195-204.

The first of two articles describing ways in which existing computer installations have been used by a university French department. A format which presented multiple choice questions in sets of five was applied to reading and grammar units. Usage recording facilities, added later, showed disappointing results, but a marked improvement was found when units were specifically linked to weekly assignments. A new type of programme designed for vocabulary review included significant improvements: individual student performance could be recorded, and no student was ever called upon to repeat an item which he had already answered correctly. Further useful features built into this programme were a record of how frequently wrong answers have been selected for each item, an automatic print-out of the entire contents of each unit, and the maintenance of a continually updated alphabetical master list of all words so far included in the total set.

35. CONNORS, Kathleen

Performance measures in L2: classification and correlations. *Bulletin of the CAAL* (Montreal), 5, 2 (1983), 117-41.

Basing research on 23 teenage Canadian anglophones in French 'late' immersion programmes, an attempt is made to discover whether the dichotomy between 'grammatical' and 'communicative' competences, or something analogous such as the degree of detail, can explain why different language measures often diverge in their evaluation of student performance.

Various text types were devised to give a four-way measure on a scale representing more or less grammaticality against more or less detail. These tests are described and justified. Thirty-one measures were eventually produced and Pearson correlation coefficients between them were established. The correlation matrix produced showed no warrant for the hypothesis that measures might be grouped according to two obvious criteria alone, but relations among different types of measure became clearer. An evaluation that was both grammatical and communicative correlated highly with measures of different types. The self-evaluation by pupils correlated with 11 measures showing that self-assessment is a useful device; rate of speech correlated with 14 other measures, but the cloze test

results correlated only with the tests of grammatical correctness, suggesting that they do not portray a global language proficiency.

36. CROSS, David

Tests of substitution: a simple alternative to cloze. *System* (Oxford), 9, 1 (1981), 51-3.

A test of substitution can be designed, administered and marked in a matter of minutes. Unlike cloze it does not assess comprehension. Like cloze it demands sensitivity to grammatical function and appropriateness of lexis. At an individual level it provides an indication of vocabulary size in the target language; it may indicate oral fluency. At group level a calculation of verb to noun ratio has proved an efficient indicator of oral proficiency, unlike cloze which provides correlation only for within-group analysis. Repeated errors indicate remedial programmes.

37. CROSS, David

Testing EFL competence. *British Journal of Language Teaching*, 22, 3 (1984), 173-4.

The EFL world would benefit from adopting graded objectives as used for setting targets and testing modern language performance in many British schools. Examples of tests based on graded objectives are quoted; on these the author has based a test battery for Arabic speakers learning English and it has attracted the attention of the Egyptian Ministry of Education. Course writers, inspectors and teachers in training will all benefit from the clear criteria of learner performance that these tests demand.

38. CROSS, David

The effects of a change from a structural to a unit-credit course. *British Journal of Language Teaching*, 23, 3 (1985), 145-51.

Two cohorts of pupils from a London comprehensive school, both consisting of top-set pupils who had chosen to continue French after the compulsory three years, were given batteries of 11 tests at the beginning and end of their fourth year. Cohort I had used the older structural *Cours Illustré* throughout; Cohort II used the newer unit-credit *Eclair* for three years, then switched to *Cours Illustré*.

It had been hypothesised that each group would score higher on a different set of tests, but the differences were much smaller than expected. At the start of the fourth year there were differences significant at the 1% level in only three of the 11 tests (Cohort II lower in substitution test, higher in fluency and number of utterances), and at the end of the year there was only one such

difference (Cohort II higher in cloze test). Fourth-year teachers appeared to have emphasised different skills with each cohort to compensate for the perceived weak points of earlier courses and thus 'iron out' differences.

39. CUMMINS, J

Psychological assessment of immigrant children: logic or intuition? *Journal of Multilingual and Multicultural Development* (Clevedon, Avon), 1, 2 (1980), 97-111.

Current assumptions regarding psychological assessment of immigrant and minority-language children are examined. Immigrant children tend to acquire fluent surface skills in their second language (L2) more rapidly than they develop L2 conceptual and literacy skills. Data are presented which show that immigrant children require, on the average, at least five years of residence in the host country to approach native norms in L2 conceptual and literacy skills. Failure by psychologists to take account of this developmental process, and of the limitations of psychological assessment instruments, can result in incorrect diagnoses of immigrant students' academic difficulties and, consequently, in inappropriate academic placement. Some of the potential pitfalls in current assessment procedures are illustrated by examples from an ongoing study in which the psychological assessments of over 400 immigrant students are being analysed, and suggestions are made for developing more appropriate assessment procedures for immigrant and minority-language children.

40. CURTIN, Constance, and others

The Pimsleur Battery as a predictor of student performance. *Modern Language Journal* (London), 67, 1 (1983), 33-40.

The predictive power of the well-established Pimsleur Language Aptitude Battery is called in question by a study of student performance at University High School, Urbana, Illinois. Final grades of students in French, German, Latin and Russian are compared with the predictions made using the battery for incoming freshmen. The population is a combined 7th-8th grade of 311 uniformly high-aptitude students.

Pearson correlation coefficients are established between grade point averages on entering and leaving the school, and the final total of the Pimsleur battery scores. Only low positive correlations are demonstrated. Correlations with the six sub-tests composing the battery are also weak, and a stepwise multiple regression analysis of the correlation matrix shows that the strongest single predictor is

the overall school ability factor, represented by the grade point average on entry, particularly for Latin. The next strongest predictor is ability to analyse language. It is concluded that no component of the Pimsleur battery is strongly predictive and that the total score predicts better, though not very reliably. Supplementary predictors or guides to student abilities are needed. However, the battery has a diagnostic value for teachers.

41. CZIKO, Gary A

Psychometric and edumetric approaches to language testing: implications and applications. *Applied Linguistics* (Oxford), 2, 1 (1981), 27-44.

A psychometric test is designed to maximise individual differences, resulting in scores which have to be compared with those of a comparison or 'norm' group; an edumetric test yields scores which are meaningful in their own right. Integrative tests (e.g. cloze and dictation) are now being used in contrast to more traditional discrete-point tests (e.g. vocabulary and grammar tests). Criticism of norm-referenced tests questions their appropriateness, since the test items are not related to the objectives of any particular programme, whereas domain- and criterion-referenced tests, which are primarily edumetric, allow scores to be interpreted not in relation to a norm group but in reference to a well defined domain of behaviour for which, in the case of criterion-referenced tests, some criterion has been set. Problems arise, however, in specifying the domains of behaviour.

The relation between the psychometric/edumetric dimension and the integrative/discrete-point dimension is considered. Integrative tests seem to be more in keeping with the edumetric approach, though they are usually put to psychometric uses, because they are direct measures of language proficiency. They could probably be adapted into tests with excellent edumetric characteristics [different combinations of variables are discussed].

42. CZIKO, Gary A

Improving the psychometric, criterion-referenced, and practical qualities of integrative language tests. *TESOL Quarterly* (Washington, DC), 16, 3 (1982), 367-79.

This research was conducted to determine whether modifications of existing integrative testing procedures could be made to improve the psychometric, criterion-referenced, and practical qualities of those tests which involve text presentation and reconstruction. For this purpose, a dictation test constructed of 14 segments varying

from 2 to 21 words in length was administered to three groups, one of foreign university students taking ESL courses, another of foreign students succeeding at graduate studies, and a third of native English-speaking students. In addition, a 'copytest' was created by presenting the same text on a computer-controlled visual display screen in a manner directly analogous to that of the dictation test. The copytest was administered to a smaller group of foreign and native English-speaking university students.

Results of the dictation test indicated that scoring the protocols by segment using an exact-spelling criterion provided an integrated measure of language proficiency with improved psychometric criterion-referenced, and practical qualities. Of particular interest was the finding that the 14 dictation segments scored in this manner formed a Guttman scale of high reproducibility and scalability. The implications of these findings as well as those of the copytest are discussed.

43. CZIKO, Gary A, and Jennifer Lin Nien-Hsuan
The construction and analysis of short scales of language proficiency: classical psychometric, latent trait, and nonparametric approaches. *TESOL Quarterly* (Washington, DC), 18, 4 (1984), 627-47.

The study reported here used classical psychometric, latent trait, and nonparametric approaches to analyse 13- and 14-item scales of English language proficiency. Tests of English listening comprehension (dictation) and reading ('copytest') were constructed by modifying the standard dictation testing procedure to create items of text segments which varied considerably in both length and difficulty. Both the dictation and copytest were found to be homogeneous, cumulative scales of language proficiency with high reliability and validity. Log ability scores provided by Rasch analysis were found to correlate better with other measures of language proficiency than did the dictation and copytest raw scores. These findings indicate that the two language testing techniques investigated provide a useful, innovative approach to measuring general aspects of language proficiency. The theoretical and practical advantages of this approach over other language proficiency measurement techniques are discussed, as are implications for measuring other cognitive variables.

44. DAVIDSON, Fred, and Grant Henning
A self-rating scale of English difficulty: Rasch scalar analysis of

items and rating categories. *Language Testing* (London), 2, 2 (1985), 164-79.

The study sought to demonstrate the utility of Rasch Model scalar analysis when applied to self-ratings of ability/difficulty associated with component skills of English as a second language. Eleven skill areas were rated for difficulty on a seven-point Likert-type scale by 228 ESL students at the University of California. Following appropriate tests of unidimensionality, both skill area items and rating categories were calibrated for difficulty, examined for fit to the Rasch Model, and plotted to provide visual representation of the nature of the item characteristic curves. Specific suggestions were made for the improvement of the rating categories of the self-rating scale, and skill areas most susceptible to self-rating error were identified. It was concluded that scalar analysis of the kind considered here is feasible with self-rating data, and that other rating scale procedures such as those employed to rate proficiency in foreign language speaking or writing would probably benefit from similar scalar analyses.

45. DAVIES, Alan

Criteria for evaluation of tests of English as a foreign language. *Views on Language and Language Teaching* (Athens), 8, 1 (1982), 7-16.

The paper makes general comments about test choice and test construction, then discusses six examples of British EFL tests currently published, widely used, and emanating from the following examination boards: RSA; Cambridge; Trinity College, Oxford; ARELS; the Panhellenic Association of Foreign Language School Owners (PALSO). They are assessed against six basic maxims and a tripartite definition of validity. The maxims assert that most linguistics is normative, language teaching samples by guesswork, language testing levels are intuitive, criterion and norm-referenced tests are the same thing, teaching needs systems of language levels which in turn need external validation. Validity¹ concerns the internal logic and interdependence of the test; validity² is the acceptance of a test as equivalent to an existing measure; while validity³ is that established statistically, subsuming concurrent and predictive validity.

Selections from the publications of the boards are reviewed and their respective merits and shortcomings assessed. All fail by the criterion of validity³ though ARELS is honourably mentioned as having at least attempted it. It is concluded that all tests have to be

assessed in terms of their purpose. They must be locally useful and globally accepted.

46. DAVIES, Alan

Validating three tests of English-language proficiency. *Language Testing* (London), 1, 1 (1984), 50-69.

Language test construction is discussed as a possible variety of general psychological test construction. Validation studies of three British English language proficiency tests are described. The process of concurrent and predictive validation for these three tests (EPTB, ELBA, ELTS) is discussed and the conclusion drawn that while these procedures are necessary they do no more than determine the accuracy of the preliminary construct and content validity consideration.

47. DAVIES, Alan

Computer-assisted language testing. *CALICO Journal* (Provo, Utah), 1, 5 (1984), 41-2 and 48.

At present CALL (computer-assisted language learning) exercises appear to be more of a testing than of a teaching type, though it is by no means certain that we know what to do. What benefits there are are limited. The three areas where CALL seems of interest in connection with testing are: (1) self access/learner-related testing in which entry points to a given population of items (e.g. a grammar test or a set of reading comprehension tests) can be swiftly determined. (2) the use of redundancy (both deletion and addition) in which a range of cloze and reading-speed tests can be called up with built-in within-text, as well as between-text, grading. Clozentropy, the group-related method of close scoring, is a special case of this type of testing. (3) Diagnostic testing (of grammar, vocabulary, common errors, etc.), combining delicate grading as well as the repeated testing of parallel item tokens of the same type. Diagnostic testing has to date been highly desired but somehow unattainable - a sort of pseudo-procedure. It does seem now that through CALL, a large enough item-bank memory, and a swift enough access can produce reliable and valid diagnostic instruments.

48. DAVIES, Alan

ESL expectations in examining: the problem of English as a Foreign Language and English as a Mother Tongue. *Language Testing* (Reading), 1, 1 (1984), 82-91.

The article makes a plea for the recognition, in testing terms, of multiculturalism in Britain today. The status and definitions of ESL

and EMT are queried in this light. Points are raised about testing in EFL (basic agreement exists on linguistic sampling and proficiency), EMT (less substantial test provision, goal definition, and no real clarification of what EMT really comprises) and ESL; in the latter case, two operational approaches are propounded. In ESL1, the particular, limited needs of ESL users can be focused upon (e.g. the JMB Test in English Overseas), whereas ESL2 involves giving a transactional, general EFL test. O-level English examinations are inappropriate for non-native speakers as they are too literary, and too culturally biased. Moreover, insufficient attention is paid to the contributions made to the development of EMT itself by speakers from minority backgrounds. The bias in favour of the majority culture is discussed. [Appendix: EMT O-level Interpretation Paper and the Test in English Overseas].

It might be appropriate to use O-levels for both EMT and ESL users, providing that discrimination was made on the basis of levels awarded; alternatively, an O-level could be granted on cumulative class work, or a range of overlapping tests could be provided.

49. DE JONG, John

Luistertoetsen moderne vreemde talen: nieuwe mogelijkheden. [Listening tests for modern languages: some new possibilities.] *Levende Talen* (The Hague), 368 (1982), 12-26.

This article describes the types of listening test currently being produced by CITO (the Central Institute for Test Development) in the Netherlands, and discusses their shortcomings. CITO have recently been experimenting with two new types of test: a short passage followed by a true/false question; and a gap-filling test, where a word or phrase is deleted from a record passage, and the testee is required to say which of two alternative answers fits the gap. Potential users seem on the whole to like these tests, and preliminary validation shows that the tests discriminate well between native speakers of English and Dutch learners of English. Some problems with the true/false tests are discussed, but CITO will be incorporating gap-filling items into their language tests in the near future.

50. DE JONG, John H A L

Listening, a single trait in first and second language learning. *Toegepaste Taalwetenschap in Artikelen* (Amsterdam), 20, 3 (1984), 66-79.

Research on language universals suggests that a single principle may underlie both first and subsequent language acquisition. A single trait also seems to underlie both native and non-native

performance on a listening comprehension test described here. The test uses live radio recordings in English cut into samples lasting about 20 seconds each and containing a deletion. Testees listen to the tape once and have to supply the deletion from two options.

The trait being measured was an ability to understand the foreign language like comparable native speakers and the hypothesis was that this trait discriminates, not only among foreign learners, but also, provided the test is not too easy for them, among native speakers. The test had 59 items and was carried out with two groups totalling 14 native speakers and one group of 575 Dutch learners of English. The results are presented and analysed both on the Rasch one-parameter latent trait model, and by classical test theory. A comparison of scores on the 40 best items and 19 rejected items is made, and different traits and relationships are revealed between the three groups tested, including factors that may be accounted for by age, intelligence and knowledge of the world. Universals and language transfer may operate at the receptive and productive level respectively, encouraging the primacy of reception in teaching.

51. DEYES, Tony

Towards an authentic 'discourse cloze'. *Applied Linguistics* (Oxford), 5, 2 (1984), 128-37.

A more discourse-oriented approach to cloze testing is needed. Units based on communicative value are more appropriate deletion items than single words if the aim is to test comprehension of a text. Ways of dealing with the identification, recoverability and relevance of such units are discussed. The degree of communicative dynamism conveyed by units can be established by criteria at several levels of analysis - syntactic, grammatical, phonological and semantic.

A prototype 'discourse cloze' test was devised by deleting theme, rheme and transition items from a text in Portuguese, using the above criteria. Roughly every seventh syntactic unit was deleted. Testees were not restricted to one-word answers, since the aim is to recover 'information' rather than linguistic items. Results showed that recoverable rhemes are those whose content can be derived by knowledge of linguistic stereotypes (collocations) and/or by stereotype knowledge about the world. In choosing communicative items to delete, the tester should ensure that they are relevant to the topic of the whole discourse. [Illustrated in an example test version.]

52. DOUGLAS, Dan, and Larry Selinker

Principles for language tests within the 'discourse domains' theory of interlanguage. *Language Testing* (London), 2, 2 (1985), 206-26.

Test takers should use a current version of their interlanguage (IL) to create personalised, prototypical contexts (discourse domains) in order to make test situations intelligible. Learners respond to test texts, for example, by either accessing existing domains, producing a temporary one, or being unable to formulate any control structure at all. Learner control of the operative domain seems to produce enhanced communicative sophistication, particularly in the area of correction and mitigation strategies.

Reference is made to research undertaken by the authors, especially in the collection of SPEAK test data [appended tapescripts], which purportedly underscores the 15 IL/discourse domain hypotheses presented in the article. The latter include suppositions about the need for valid tests to engage relevant learner domains if they are to provide genuine assessment, the seeming linkage between the differential creation of IL structures and the realities of particular situations, and the idea that training in test-taking strategies can actually help a testee to deploy a misleading 'meta-domain' which might vitiate interpretation of his/her test performance. These and other connections between testing and SLA/IL research provide a richer conceptualisation of the language testing process than has heretofore been possible.

53. DRYDEN, Norman H

Written text comprehension speed and accuracy: a response recording and timing system using a PDP11/40 computer. *Work in Progress* (Edinburgh), 15 (1982), 19-25.

A technique is described for presenting informants with written text, and testing the speed and accuracy of their comprehension. The aim was to present a series of sentences, each of which was split into three phrases. Each sentence was to be followed by a relevant question to elicit a yes/no response. Timing was to be under the control of the informant, with each phrase or question replacing the previous one, so that there could be no referring back. The experiment was implemented on a PDP11/40 DECLAB computer, using a video-display unit incorporating a graphics mode for the presentation of text. The time parameters were measured using the LPSKW real-time clock, which is programmable. [Programming method.] The results for each informant were held as a file on the computer disk for later listing and analysis. The whole procedure worked reliably when used by informants and was used to test comprehension of written German by native and non-native readers.

54. DUCROQUET, Lucille

Objective or subjective testing? *System* (Oxford), 8, 3 (1980), 237-43.

The claim for objectivity has been commonly used in the past few decades as a criterion for excellence in language tests. This claim is examined here, firstly through a critical presentation of the characteristics of objective tests, and secondly as compared with integrative tests (with particular reference to John Oller Jr., 1979). This examination leads to a discussion of what the best approach to testing language might be, and it is argued that a combination of integrative and objective techniques should ensure maximum validity and reliability.

55. DÜWELL, Henning

Leistungsbereiche der Lernerfolgsüberprüfung. Eine Untersuchung zu den Aufgabenstellungen von Klassenarbeiten des Englisch- und Französischunterrichts in Klasse 10. [Applications of achievement testing. A study of tasks in class tests in English and French in class 10.] *Die neueren Sprachen* (Frankfurt am Main), 29, 5 (1980), 446-63.

A study of 654 French and English class tests written by 10th grade pupils in 35 schools of various types in the Federal Republic of Germany is discussed. The tests are examined with regard to (1) their points of execution in the temporal/organisational frameworks of their respective language classes, (2) their content, and (3) their different types of item. A total of 1,448 required linguistic operations are located and categorised according to a descriptive model as either isolated language elements or complex language applications. The study is designed to describe language achievement testing (Lernerfolgsüberprüfung) practices in the 10th grade. The required linguistic operations are examined to determine their usefulness as testing criteria for the 16+ school-leaving qualifications currently under discussion.

56. ELLIS, Rod

Communication strategies and the evaluation of communicative performance. *ELT Journal* (London), 38, 1 (1984), 39-44.

In evaluating communicative performance in a second language, attention should be paid not to correctness, intelligibility or style but to communication strategies. The article focuses on two general strategies, avoidance and paraphrase. The possibilities are that 'there are quantitative differences between native speaker and L2 use of communication strategies (L2 speakers will employ strategies

more frequently), and that native speakers employ more achievement strategies' such as paraphrase, and fewer 'reduction strategies' such as avoidance (qualitative differences). The use of strategies varies according to the proficiency level of the learner, less able students preferring an avoidance strategy, the more able preferring paraphrase.

A study was carried out to see whether these differences are matched by differences between L2 learners and comparable native speakers. Subjects were young L2 learners aged 10-12 and native speakers of the same age. The children had to tell a story from pictures. Various key 'information bits' were selected which should have been included. It was found that the L1 children encoded substantially more information about the pictures than the L2 children and resorted less to avoidance and paraphrase strategies. If teachers attend to the degree to which learners avoid reference to important items and paraphrase information they do decide to encode, they may be able to form a fairly reliable assessment of learners' performance. They can do this impressionistically or by using a chart.

57. ELLIS, Richard J

The assessment of oral skills. *Teaching English* (Edinburgh), 14, 3 (1981), 19-23.

An oral English assessment (in mother-tongue teaching) should (a) signal to the pupil where progress has been made and where more work needs to be done, and (b) signal to the teacher where more emphasis should be placed in preparing the pupils. Schools should devise their own assessment of the kinds of skills they think worthwhile for their pupils: a 'package' of skills which will help any pupil in various vocational and social situations. Teachers should all be using the same terminology for the same things. The use of a profile sheet is recommended on which the various skills thought necessary are indicated as segments of a circle. As pupils progress through secondary school, their teachers will shade in the various segments until they are complete. Such a chart is more suitable for a slow process like acquiring oral skills than a one-off test; progress and shortcomings can be seen at a glance.

58. ERVIN, Gerard L

Assessing oral proficiency: some practical considerations. *Russian Language Journal* (Michigan), 34, 118 (1980), 7-14.

Proficiency tests measure the student's performance against the subject matter, rather than against a particular syllabus. They measure what the student can do in the language. The FSI, or Civil Service, oral interview technique gives the interviewer a good idea of the student's ability. Examiners must be trained in interviewing techniques and assigning ratings. At least one of a pair of examiners working together must have native or near-native competence in the target language. Ratings on the proficiency scale represent a range, not a point.

The academic community should consider the adoption of a common standard, so that meaningful comparisons can be made between people with different trainings. Proficiency exams will probably never replace traditional achievement tests because of the time and cost involved, but they do not need to be given so frequently. The adoption of a widely known standard would be helpful to employers, language schools and academics.

59. ESSER, Ulrich

Ein allgemeinspsychologischer Ansatz zur Diagnose der Fremdsprachenbefähigung. [A general psychological approach to foreign language learning ability.] *Deutsch als Fremdsprache* (Leipzig, GDR), 19, 2 (1982), 76-81.

This article outlines a test to establish degree of language learning ability and gives first results of the test relating to objectivity, reliability and validity. The test is non-verbal, based on school-level language training, and, as far as possible, culturally impartial. The aspects of language-learning ability tested are: associative learning in pairs and the ability to integrate information, learn rules and form analogies. The theory underlying the development of the test is that aptitude tests to date cover aspects of language learning ability which have merely been deduced and which themselves represent functions of general cognitive conditions for human information processing and require examination and diagnostic analysis.

60. EUBANKS, Frances

Oral testing in the examination context: the short talk and its summary. *Perspectives* (Rome), 8, 1 (1984), 31-45.

Testing in Italian universities is done by public oral interview; here a test is described as developed for the end of the second year of a four-year English programme in a department of Business and Economics, at intermediate level. Candidates were asked to prepare a five-minute talk on a topic of their choice, but notes and written texts were not allowed during the examination. The talk would be

presented to a student partner chosen on the spot, with the examiner present and assessing the communication between the two students. The partner would listen and then give a summary of the talk back to the first student. Then the roles would be reversed.

Performance was assessed by more than one examiner on three five-point scales relating to Quality of Topic, of Understanding and of Summary, and the main criterion was ability to organise discourse as evidenced by clarity and use of detail. This test format produced more language under student control than most oral tests, while at the same time inducing more relaxation, spontaneity and interest in the candidate. It also met most established criteria for communicative testing. Oral examiners should limit their role to that of assessors in view of the success achieved by non-participation in the dialogue.

61. FARHADY, Hossein

Measures of language proficiency from the learner's perspective. *TESOL Quarterly* (Washington, DC), 16, 1 (1982), 51-9.

This paper examines the importance of learner characteristics in relationship to learner performance on ESL tests. Test taker characteristics are not included in the design of most ESL tests. Empirical evidence is provided to support the hypothesis that performance on various ESL tests is closely related to test takers' educational and language backgrounds. In order to account for these factors and thus decrease test bias, the theoretical definition of language proficiency should be modified. Some guidelines to dealing with test taker characteristics are suggested concerning (1) psychometric factors (improving the correspondence between test items and instructional objectives), (2) typologic factors (different types of test need clearer item specification, and programmes need clearer instructional objectives), and (3) learner factors - adjustment formulas are needed in the short term to account for learner variables, and in the long term, analysis of learner needs, establishment of instructional objectives, then a two-phase testing programme to assess, first, 'base level' English, then proficiency in various academic areas.

62. FISCHER, Robert A

Measuring linguistic competence in a foreign language. *IRAL* (Heidelberg), 19, 3 (1981), 207-17.

The article evaluates several testing techniques which provide a relatively direct assessment of the student's linguistic competence, and reports the findings of pilot tests carried out in first-year

university French classes. It was concluded that the most direct means of evaluating linguistic competence seems to reside in multiple-choice test items that require the student to make a comparative judgment of the grammatical acceptability of the options given to him. The nature of the response mode limits its utility to the early phases of classroom testing. Extended use of multiple-choice as a testing device creates a passivity syndrome and yields only marginal returns. After the initial stages of language instruction, the cloze procedure holds the greatest promise as a method for assessing linguistic competence because of its flexibility in deletion patterns and variety of examination formats. In tests of function words the basic open-ended form presents a sufficiently challenging task to the student and provides reliable results. In tests of inflected content words either the lexical inventory form or the open-ended form, which requires more active recall, produces an economical testing device that accurately measures both morpho-syntactic and semantic knowledge. The cloze procedure not only provides a relatively direct evaluation of the student's linguistic competence but also introduces him to the contextual demands of a communicative situation. As such it leads the student from purely linguistic competence to communicative competence.

63. FISCHER, William B

Not just lip service: systematic oral testing in a first-year college German program. *Die Unterrichtspraxis* (Philadelphia, Pa), 17, 2 (1984), 225-39.

For five years Portland State University has been carrying out oral tests every two weeks with students on introductory German courses, using the ACTFL/ETS Provisional Proficiency Guidelines and the Interagency Language Roundtable interview model of oral proficiency testing, in a modified form, to establish the floor level of students as they pass through the course. Regular tests of listening and reading comprehension will subsequently be introduced. Oral examiners are trained to give short and sympathetic interviews. Advice on oral examining is given based on their experience. The test content and procedures are described and illustrated. Results tend to confirm the value of the ILR interview model and diminish the paperwork involved in other forms of testing.

64. FOLEY, Joseph A

More questions on assumptions about cloze testing. *RELIC Journal* (Singapore), 14, 1 (1983), 57-69.

The author takes up the suggestion made by Johnson (1981) that some of the assumptions made about cloze tests may be misleading. Research which has now been going on for 30 years indicates that cloze procedure is a criterion-referenced test which will identify whether an individual or a group possesses the basic skills in the target language. Cloze is a very blunt instrument for testing but within its limits it is valid and reliable. However, changing the variables such as deletion rate, scoring procedure or using a different text may well result in a radically different test each time. The raw data obtained from a cloze test cannot be used as continuous data, such as would be derived from most psychometric-type tests, but as an indicator of one of three levels: frustrational, instructional and independent. Guidelines are given for the preparation of a cloze test.

65. FOLEY, J A

A rationale for the cloze procedure. *ITL* (Louvain), 64 (1984), 71-89.

This article attempts to answer two questions: first, the question of what cloze actually is. Explanations using the principle of 'redundancy' in the text and the 'grammar expectancy' of the decoder seem to be more satisfactory than theories based on gestalt psychology or 'information theory'.

Secondly, the question of what cloze actually does. This would appear to be only satisfactorily answered by reference to the syntactic, semantic and rhetorical elements of language. Comprehension assessment for which cloze is often used is normally understood as getting at the conceptual world of the encoder. This cannot be directly tapped by cloze as it would involve skills over and above those accessible by cloze, such as skimming, scanning, reorganisation, inference and evaluation.

66. FORD, Cecilia E

The influence of speech variety on teachers' evaluation of students with comparable academic ability. *TESOL Quarterly* (Washington, DC), 18, 1 (1984), 25-40.

This study considers teachers' evaluations of students who perform equivalently in written academic work but who have distinct speech varieties. Stimulus materials to which teachers were asked to react were collected from third and fourth graders in schools in the Southwest of the United States. The materials included speech samples (some of which exhibited Spanish-influenced English and some of which did not) and writing samples which exhibited no Spanish influence. Forty teachers evaluated paired speech and writing samples, assuming that each pair had been produced by one and the same student. In fact, the writing samples had been

independently evaluated as equal before they were paired with the speech samples, half of which exhibited Spanish influence.

In their evaluations, all teachers favoured the non-Spanish-influenced speakers despite the evidence of comparable written work. Length of teaching experience and teacher ethnicity did not significantly affect the degree of directionality of differences in the teachers' rating of the speech groups. However, in evaluating social status, the teachers who spoke Spanish as a first language showed significantly less difference between their ratings of the two groups than did the other teachers.

This study offers evidence that, due to the stereotyped expectations of teachers, Spanish-influenced English speakers are at a disadvantage in schools in geographic areas of the United States where there are large Hispanic populations. This evidence is also of relevance to other parts of the world where linguistic prejudice may exist. The findings further suggest that a teacher's first language may affect that teacher's perception of students who speak distinct speech varieties.

67. FOULY, Kamal A, and Gary A Cziko

Determining the reliability, validity, and scalability of the graduated dictation test. *Language Learning* (Ann Arbor, Mich), 35, 4 (1985), 555-66.

The purpose of this study was to examine the reliability, validity, and scalability of a new technique for constructing, administering, and scoring dictation tests to measure second language proficiency. The 'graduated' dictation test used in this study involved the auditory presentation of progressively longer text segments to be written down by the examinees. A total of 624 students learning English as a second language at an American university took the dictation test and several other measures of English-language proficiency. The results of this study demonstrate that the graduated dictation task is a reliable and valid language testing technique. Using Mokken's criteria of scalability, the test was also shown to constitute a valid cumulative scale with no non-scale items. These findings and the advantages of this technique over commonly used dictation tests are discussed.

68. FOX, Barry

Communicative competence and the APU. *Language for Learning* (Exeter), 1, 3 (1979), 159-67.

ABSTRACTS

In 1978 the Assessment of Performance Unit published two pamphlets setting out its theory of language performance and criteria for assessment. Since the APU is planning to assess the language of schoolchildren in the UK on an unprecedented scale and in a way that will exert a powerful influence upon practice in the schools, it is interesting to note how the APU deals with the question of communicative competence. What emerges from the two documents is that, while the APU shows theoretical understanding of some aspects of communicative competence, the methods of assessment it has adopted in practice are arbitrary and inadequate to the task which it has set itself. 'Skill-testing' would be a better term since that does not suggest an all-round ability but rather a number of possibly unrelated and minor abilities. The APU provides no theoretical basis for believing that those areas which it wishes to test seriously affect communication, and since the areas chosen correspond to what is traditionally taught in schools, with all that that implies, one is led to conclude that it is not patently interested in communicative competence as such, a view reinforced by the fact that nearly all the 'errors' given as examples do not interfere with communication. It is clear that the APU is mainly concerned with assessing the politer skills of the politer linguistic activity of writing. Hymes warns and Labov demonstrates that communicative competence is not realised only in standard forms of language.

69. FRANCIS, J C

The reliability of two methods of marking oral tests in modern language examinations. *British Journal of Language Teaching* (Birmingham), 19, 1 (1981), 15-23.

The efforts made by one GCE board to improve the reliability of oral tests are described. Multiple marking would doubtless be desirable to increase reliability but the cost of training the necessary examiners would be prohibitive. Some analytic methods can also increase reliability. The board in question offers a one-day training session in examining techniques and a one-day standardisation meeting for all newly appointed examiners.

Until 1976 all oral tests conducted by this board used a global impression technique. An experimental analytic scheme which gave different weightings to different aspects of oral ability did not prove conclusively superior so a further investigation was designed in which the oral tests of about 100 candidates in Italian were double marked on three separate occasions. The marking of candidates in a live oral test appear to be more reliable than marking taped tests. Results do not show any real difference in the reliability of the

analytic method as against the global impression technique. However, all the examiners preferred using the grid system, which is analytic, although not all were equally effective using it, and some were very reliable using the impression method.

70. FRETWURST, Peter

Ständige Leistungskontrolle - eine unabdingbare Komponente des Unterrichtsprozesses. [Continuous assessment at an indispensable part of the teaching process.] *Fremdsprachenunterricht* (Berlin), 24, 6 (1980), 257-60.

Continuous assessment is important in foreign-language study both to motivate the pupils and to pinpoint gaps in their knowledge. The teacher is then able to adapt future lessons and to fill these gaps in accordance with the specific needs of the class. In order to reflect progress accurately, all the various activities which form part of language teaching must be examined. Accent and intonation must be tested as well as grammar and vocabulary. Tests should not only be concerned with what has been covered recently but also with what has been dealt with in the more distant past.

Teachers should make sure that the pupils grasp the long-term aims of the assessment system so that they develop a positive attitude to it. They should be made aware of the criteria by which their work is judged. Indeed, pupils should be encouraged to assess their own progress and to identify the areas of knowledge where improvement is called for.

71. FRETWURST, Peter

Die Bewertung der Leistungen als notwendiger Bestandteil des pädagogischen Prozesses im Fremdsprachenunterricht. [Assessment of pupils' performance as an essential part of the education process in foreign language teaching.] *Fremdsprachenunterricht* (Berlin), 24, 7/8 (1980), 322-6.

The wider goals of a communist education and the effects on the character development of individual pupils should determine the method of assessing education standards. Tests ought to be planned at the same time as the lessons rather than afterwards. It is the teacher's responsibility to see that the marking system is fair and consistent. However, the teacher should comment on performance according to the individual requirements of each pupil. A pupil's attitude to a subject is greatly influenced by the teacher's reaction to his performance. Whenever possible each pupil should be made aware that he is making progress.

Assessing grammatical exercises, translations and the like is comparatively straightforward; the more creative use of a foreign language poses a different problem. A teacher must assess to what extent mistakes hinder effective communication while bearing in mind that if too much importance is attached to correctness, the pupils may be discouraged from attempting to communicate. A pupil's final course mark should be taken as his ultimate grade because otherwise he may be held back by an earlier failure which has since been made good.

72. FRIEL, Michael

The Discourse Co-operation Test. *System* (Oxford), 12, 3 (1984), 251-62.

The Discourse Co-operation Test (DisCoTest) relies on the maxims of successful conversation laid down by Grice as his Co-operative Principle and on the maxims of 'antecedence' and 'lexical specificity' derived by others from his work.

The test is a series of oral stimuli sequenced so as to be unpredictable to the testee but constant across testees. Five parallel dialogues all deriving from the same opening remark are drawn up and illustrated. They each consist of five exchanges between tester and testee and each time the tester speaks he flouts one of the above maxims, forcing the testee to implicate successfully in order to carry on the conversation. Thus, using multiple choice procedure, it probes 'the indirect utilisation of utterances' in genuine communication. It has been successfully used in Yemen and gave K-R reliability coefficients of 0.82 and 0.81. It is well suited to the language laboratory and to training in conversational awareness but needs further validity and reliability studies.

73. GAIES, Stephen J

T-unit analysis in second-language research: applications, problems and limitations. *TESOL Quarterly* (Washington, DC), 14, 1 (1980), 53-60.

A critical review of arguments for and against T-unit analysis in second-language research, the T-unit being 'a main clause plus all subordinate clauses and non-clausal structures attached to or embedded in it' (Hunt, 1970). T-unit length increases with the development of syntactic maturity. It provides a global measure of linguistic development and allows for the meaningful numerical comparisons between first- and second-language acquisition. Suggestions are made about the validity, usefulness and limitations of this and other measures of overall syntactic complexity.

74. GARDNER, R C, and P C Smythe

On the development of the attitude/motivation test battery. *Canadian Modern Language Review* (Toronto), 37, 3 (1981), 510-25.

The development of an attitude/motivation test battery is described, including, in some detail, both the rationale and techniques employed in the construction and validation of the battery as well as the issue of test reliability. By focusing on the necessary steps involved in test construction, validation and evaluation it is hoped that researchers and educators in this area will be encouraged to pay more attention to the measurement properties of their test instruments.

75. GREANEY, Vincent

Developing national attainment measures: a study of Irish experience. *International Review of Education* (Hamburg), 26, 1 (1980), 3-16.

Recent experience of test development in Ireland, a country which had no tradition of formal standardised testing, is drawn on, in order to highlight a number of practical and measurement problems that future test developers in other countries may experience. The paper focuses on a number of issues that have received relatively little attention in the literature. The practical problems which are examined include the identification of the objectives of the school curriculum, the production of test materials and the selection of appropriate samples of schools for standardisation purposes. The measurement or psychometric problems relate to curricular and statistical characteristics of individual test items, a test ceiling effect caused by some pupils answering all or almost all of the test items correctly and the lack of test sophistication among pupils and teachers. A number of conventional procedures for developing norm-referenced measures of pupil attainment may prove somewhat less than successful in countries or educational systems with common curricula.

76. GREEN, Peter S, and Karlheinz Hecht

Reliability assessment of written communicative skills. *Finlance* (Jyväskylä, Finland), 3 (1984), 1-23.

Reporting a joint research project in the U.K. and Germany conducted by the universities of York and Munich respectively, the authors show how an essay-type task - writing a letter of elicitation between penfriends - was marked differentially by teams of German and British markers. Sources of the low inter-marker reliability

were: (a) disagreement as to what constituted an error, (b) disagreement as to the gravity of errors agreed upon, (c) the penalising of pseudo-errors by German markers, (d) the heavier penalisation of grammatical errors by German markers, (e) the heavier penalisation of vocabulary and style errors by British markers.

Re-marking using a more explicit marking scheme improved intermarker-reliability, but it is suggested that a radical improvement could only be secured if the ideal target behaviour for candidates is redefined as what would be acceptable native-speaker performance among pupils of the same age and ability in equivalent U.K. schools, not that of an educated adult native-speaker. The methodology and results are described in detail, with analysis tables and discussion

77. GRENIER-HENRIE, Anne-Marie, and others

La mesure de la compétence de communication à la Direction Générale de la Formation Linguistique: les tests de performance en communication orale. [Testing competence in oral communication.] *Medium* (Ottawa), 7, 2 (1982), 113-33.

The Direction Générale de la Formation Linguistique aims to test the communicative competence of English-speaking learners of French with regard to: intelligibility, communicative strategy adopted, comprehension, grammar, fluency, and the sociolinguistic aspect. The test lasts about 15 minutes, consists of an interview and a language task (examinees are given a choice), and is recorded on tape. The examiner, who should not be the candidate's teacher, is asked to make the conversation as lifelike and as far removed from the artificial examination situation as possible.

The guidelines for the examiner are included and the system of marking explained. [Specimen language task.] The article is a continuation of a previous article on the theory of language testing (Monfils).

78. GRUNDIN, Hans U, and others

Cloze procedure and comprehension: an exploratory study across three languages. *Journal of Research in Reading* (Leeds), 4, 2 (1981), 104-22.

This cross-cultural study of cloze procedure and comprehension involved samples of 10- to 11-year-old schoolchildren in Canada, Japan, Sweden and the United States. The aim of the study was to explore the nature of what might be called 'cloze comprehension' in relation to overall or 'global comprehension' of a passage; in particular to establish (a) whether cloze procedure measures the

same facets of comprehension regardless of what language is being read; and (b) to what extent cloze procedure, in different linguistic areas, measures 'global comprehension', or comprehension of the general ideas contained in a passage, as distinct from literal comprehension.

The results of the study indicate that cloze procedure is a valid and reliable measure of certain aspects of reading comprehension in all the linguistic and cultural areas sampled. Furthermore, comprehension as measured by cloze procedure seems to be a necessary, albeit not sufficient, condition for overall or global understanding of the meaning of a passage. The study also shows that the ability measured by cloze procedure is more generalised (i.e., less text specific) than the ability measured by our global comprehension task. One implication of this seems to be that the higher-order skills necessary for global understanding do not always develop automatically once children have mastered the skills necessary for literal comprehension of simple texts. On the contrary, the higher-order skills may have to be taught systematically at an appropriate stage in the children's reading development.

79. HALE, Gordon A, and others

Effects of test disclosure on performance on the test of English as a foreign language. *Language Learning* (Ann Arbor, Mich), 33, 4 (1983), 449-64.

Specifically constructed 'disclosed TOEFL forms' were made available to students in intensive English language programmes. Later the students were administered a special TOEFL consisting of items from those disclosed forms as well as a TOEFL consisting of items to which they had not been exposed. Scores on the former test were found to be greater than those on the latter, indicating a 'disclosure effect' - i.e., showing that the students studied and recalled specific items from the disclosed forms.

Some items in the disclosed forms were to be discussed in class by the students' language instructors before the special TOEFL, and a relatively large disclosure effect was observed for these items. Nevertheless, a significant disclosure effect was also observed for items not discussed in class, indicating that the students studied the disclosed forms on their own initiative.

While some students were given six disclosed forms to study, others were given 12 forms, and the disclosure effect was found to be more pronounced for the former group than the latter. Apparently, if students must cover a relatively large number of forms in order to be exposed to all items that will appear on a later test, they will be less

likely to benefit from disclosure than if they need cover a smaller number of forms.

80. HAMEYER, Klaus

Testing oral proficiency via elicited imitation. *Revue de Phonétique Appliquée* (Mons, Belgium), 53 (1980), 11-24.

There is a need for an index to gauge the communicative competence of adults in a second language. Elicited imitation has been used in first-language acquisition as well as with the mentally retarded, but only recently has it been applied to second-language learning. Presented here is a test for oral proficiency in the imitation of German sentences. Sixteen adult learners of German with different native languages and different proficiencies in German were tested. They were presented 129 German sentences of varying length and syntactic complexity, and their imitation of these sentences was recorded for later analysis.

After the short-lived 'acoustic memory' has been overtaken by longer sentences, it is seen that the length of the correctly repeated sentences strongly correlates with the number of syntactic and semantic errors, and they in turn with the amount of German study. None of the various syntactic units is as well suited as the sentence length to measure oral competence in a simple and objective manner.

81. HANANIA, Edith, and May Shikhani

Interrelationships among three tests of language proficiency: standardised ESL, cloze, and writing. *TESOL Quarterly* (Washington, DC), 20, 1 (1986), 97-109.

Cloze, which combines the advantages of integrative testing and objective scoring, was investigated as a supplement to a standardised ESL test and as an alternative to a written composition test. Three tests of English language proficiency were given to a large group of students applying for admission to the American University of Beirut (AUB): the AUB English Test, a cloze test, and a written composition. The tests were taken by the same group of students at the same examination session, enabling direct study of interrelationships among the three measures. Regression analysis of the test scores showed that pairwise correlations were all high and that a combination of cloze test scores and AUB English Test scores significantly improved the prediction of communicative language proficiency, as measured by the composition test scores. In addition, there was a substantial residual correlation between the cloze and writing tests, which suggests that these tests may measure in common some aspects of language ability beyond those that they

share with the AUB English Test, a standardised ESL test. These results indicate that a cloze component can serve as a valuable supplement in language proficiency testing. Further implications of the findings of the study are discussed.

82. HARRISON, Andrew

Student-centred testing: assessing communication in progress. *Levende Talen* (The Hague), 372 (1982), 401-10.

Teaching procedures and assessment procedures depend on each other. A coherent method of language teaching makes assumptions about a theory of learning and these are reflected in its test procedures. A communicative approach to teaching and learning assumes that the most important principle for deciding what should be learnt is the use of language in realistic contexts. But a test is always to some extent a simulation. There needs to be enough control over the content for necessary comparisons to be made. The assessment of communication starts with an information gap which needs to be bridged - a puzzle strategy is a possible alternative for the early stages of language learning. The kind of problem which can be set for assessing communication has to be suitable for solving with language, the right level of language use, the right interest for students, relevant to language needs and generalisable to other situations. The selection of means for any assessment is guided by the following criteria: how to set limits to the permissible language (translation, closed set of responses), appropriateness, tolerance of error (until it prevents communication), the analysis of minimum essentials of communication (based on competent performances by fluent speakers). Marking could be done on the basis of credit for 'communication' plus credit for 'correctness' when communication has been achieved. The performance of each student can then build up into a series of achievements, interlocked in a sequence but not necessarily interdependent. This kind of assessment system has considerable flexibility; it can be applied to different contents and used for different purposes.

83. HENNING, Grant, and others

Comprehensive assessment of language proficiency and achievement among learners of English as a foreign language. *TESOL Quarterly* (Washington, DC), 15, 4 (1981), 457-66.

Seven subscales of English language proficiency were devised and administered to 184 Egyptian students at the third year secondary stage. The purpose was to identify optimal combinations and weights of these subscales for predicting EFL proficiency while

following constraints imposed by the nature of the syllabus. Error Identification, Grammar Accuracy, Vocabulary, and Composition subscales were indicated as best predictors, but Error Identification was ultimately replaced by Reading Comprehension to better evaluate course objectives. Error Identification was indicated as an indirect measure of composition writing ability.

84. HENNING, Grant

Oral proficiency testing: comparative validities of interview, imitation, and completion methods. *Language Learning* (Ann Arbor, Mich), 33, 3 (1983), 315-32.

Employing an initial sample of 143 adult Egyptian learners of English as a foreign language, the three oral testing methodologies of imitation, completion, and interview were compared for reliability and validity. Similarly, five components under each method, namely, raw score, fluency, pronunciation, grammar, and combined fluency-pronunciation-grammar ratings, were analysed separately and in tandem. Multi-component-multimethod convergent and discriminant validities were determined. Stepwise multiple regression was computed using FSI-like interview scores as the dependent variable. And Rasch latent trait calibration and tests-of-fit validity were computed for imitation and completion tests.

Results indicated that the pronunciation component of the imitation method exhibited highest overall validity across all indexes. The FSI-like component of the interview method ranked second and the fluency component of the imitation method ranked third. Comparison of the three oral testing methods across all components for all empirical validity indexes showed (1) imitation, (2) interview, and (3) completion methods to rank in that respective order in terms of available composite validity indexes. Regression analysis showed the FSI-like interview to be primarily related to grammar skill from among 11 independent predictors examined. Advantages of Rasch measurements were noted for management and analysis of item data.

85. HENNING, Grant

Listening recall - a listening comprehension test for low-proficiency learners. *System* (Oxford), 11, 3 (1983), 287-93.

Most widely used foreign language listening tests use a variety of multiple-choice-type test responses. Typically these tests involve listening to a stretch of narrative material and then answering questions over the content or listening to one or more voices making statements or asking questions and then choosing an appropriate

response to what has been said. Such tests are too difficult and do not discriminate among low-proficiency learners. This paper describes a listening measure which has proved to have good validity and reliability, discriminates over a wide range of proficiency, and is easy to construct and mark. The test is called 'Listening Recall', and is a type of listening cloze procedure, without random deletion. A narrative passage is prepared with a certain number of content words deleted. Subjects are given a written version to read and then they listen to the passage and try to replace the deleted words or provide words which mean the same thing. Statistical results are analysed, and the Listening Recall measure is discussed in comparison with other types of listening tests.

86. HENNING, Grant

Advantages of latent trait measurement in language testing. *Language Testing* (London), 1, 2 (1984), 123-33.

The practical advantages of latent trait measurement techniques, in this case centring on the Rasch one-parameter linguistic model, are discussed in relation to the alleged shortcomings of classical measurement theory, which is primarily concerned with identifying difficulty, variance, discriminability and validity. A summary of classical and latent trait methods is provided; classical approaches evince problems in generalising test scores beyond the sample of persons tested and the item corpus included in the test.

Both approaches are used to scrutinise a 48-item ESL reading test, which is in turn part of a diagnostic battery administered at North Texas State University [tabularised data]. The Rasch model was chosen instead of either the Birnbaum two or three parameter models because it can be used with samples of 100 or less, and required no computer data processing. The results seem to show that latent trait measurement is a more powerful analytical tool than any classical method, even identifying, for example, those candidates for whom test scores may be invalid. Advantages for test developers are enumerated, including the facilitation of alternative test construction.

87. HENNING, Grant, and others

Item response theory and the assumption of unidimensionality for language tests. *Language Testing* (London), 2, 2 (1985), 141-54.

Considerable controversy has arisen around the assumption of unidimensionality underlying the application of latent trait models of measurement. The intent of this paper is to provide a clearer

articulation of the unidimensionality assumption and to investigate the robustness and applicability of a particular unidimensional model, the Rasch Model, for use with language proficiency tests that consist of batteries of subtests in a variety of skill areas and that are applied in the testing of the abilities of students from diverse educational, linguistic and cultural backgrounds. Results of the analysis of response data from the administration of a 150-item, five-subskill ESL proficiency/placement examination to 312 entering university students indicated that unidimensionality constraints were not violated.

88. HERAS, R, and M Regan

A SGAV approach to individualisation, autonomy and contextualisation through macrosimulation. *Revue de Phonétique Appliquée* (Paris), 73/5 (1985), 95-116.

A programme developed at an Adult Migrant Education Centre is described. The aim was to use macrosimulation to encourage learner autonomy and greater individualisation within an authentic Australian context. The needs of the learners were assessed by means of a questionnaire, group discussion and informal interviews. The prime motivation was to continue education or to get a job. A list of specific functional objectives was compiled.

A television series 'A Country Practice' about an Australian family with a Down's Syndrome child was used as prior input, together with a current affairs programme on the same theme. A broad range of authentic documents and 'props' was collected. The students chose an Australian town for the setting of their simulation and created their own roles, depending on their interests and motivations. As the simulation progressed, the students gained in confidence. The triggering activity to begin the simulation was the election of a mayor. Residents kept a diary to record the previous day's happenings, and to express their feelings about them. The diaries were useful feedback for the teachers and showed the learners what language tasks they had learned. Regular times were set aside for correction and remediation, including the use of video. By the end of the course, all the students felt more confident in being able to meet their long-term goals. Macrosimulation was felt to offer greater opportunities for individualisation than any other methods. It encourages learner autonomy and self-evaluation.

89. HOLEC, Henri

Plaidoyer pour l'auto-évaluation. [The case for self-assessment.] *Français dans le Monde* (Paris), 165 (1981), 15-23.

Assessment is an essential component of the learning process. The great majority of learners already practise an informal kind of self-assessment but need to acquire the skills necessary to observe, analyse and evaluate their linguistic performance. There are also psychological obstacles to students making their own assessment of their progress, which can only be overcome by changes in attitudes to teaching and learning and to the role of the teacher. Self-assessment could constitute a potential fruitful source of innovation and renewal for language teaching.

90. HOMBURG, Taco Justus

Holistic evaluation of ESL compositions: can it be validated objectively? *TESOL Quarterly* (Washington, DC), 18, 1, (1984), 87-103.

The purpose of this study is to capture what it is that trained ESL composition readers do when evaluating essays. Using a sample from 3,512 compositions written as part of the Michigan Test of English Language proficiency, the holistic judgements of markers are compared with results obtained on a variety of objective measures: (a) of length, (b) of syntactic complexity and (c) of error. Three bands of results are compared around the middle of the marking scale and tables of descriptive statistics given. Some measures distinguish equally across all three grades, but most tend to associate two of the grades closely together. To determine the relationship between them, three statistical analyses were run on the data, namely: one-way analysis of variance; orthogonal comparisons; stepwise discriminant analysis. Results show that 84 per cent of the variance among the three bands was accounted for by five measures: second-degree errors per T-unit; words per sentence; dependent clauses, co-ordinating conjunctions, and error-free T-units per composition. Readers seem to apply these measures on a funnel model using gross estimates first and then narrowing down to finer distinctions. Holistic evaluation by trained readers is shown to be adequately reliable and valid.

91. HUDSON, Thom, and Brian Lynch

A criterion-referenced measurement approach to ESL achievement testing. *Language Testing* (London), 1, 2 (1984), 171-201.

The authors discuss the divergent merits of norm-referenced (NRM) versus criterion-referenced measurement (CRM) against the specific background of the final examinations in the UCLA ESL service courses. Whereas NRM highlights student differences,

giving a broad, fairly imprecise view of relative standing, CRM adduces a more descriptive picture, measuring absolute standing against defined instructional goals or sets of skills. CRM, particularly in achievement testing contexts, allows more explicit inferences to be made about the success of teaching programmes. Details of the UCLA testing situation and the attendant constraints are followed by general consideration of CRM/NRM techniques, with reference to item, reliability, and validity analysis. Development of the UCLA test battery is described, as is the influence of the university's Testing Project, of which this article is a product.

Subsequent CRM study [tabular data] identified items which discriminated well between masters/non-masters of particular linguistic skills, suggested possible methods of ensuring test consistency, and underscored a more direct feedback relationship between testing and teaching than could be identified by using NRM. The research conclusions are only tentative, however.

92. HUGHES, Arthur

Conversational cloze. *Studi Italiani di Linguistica Teorica ed Applicata* (Rome), 9, 2 (1980), 167-78.

Cloze tests are usually used to measure overall ability in a second or foreign language. Research is described in their use to measure a particular language skill, oral ability, testing the hypothesis that a cloze test based on a recorded conversation between native speakers of English would be superior to a cloze test based on a prose passage as a predictor of EFL students' ability to take part in conversation or discussion in English. Validation was against teachers' ratings of their students in this skill.

A stretch of conversation was recorded and transcribed and every seventh word deleted. There were obvious differences from formal prose style - contractions, hesitations, references to familiar institutions. Subjects were 64 overseas students of various linguistic backgrounds on pre-sessional English courses at the University of Reading. Both the conversational and the prose cloze test were scored by the exact word method. Native speakers' responses to the conversational cloze were used as the basis of decisions on acceptability of responses. Results showed that conversational cloze was consistently superior to prose cloze as a measure of oral ability.

93. INGENKAMP, K, and B Wolf

Research in oral secondary-school leaving examinations in Germany. *British Journal of Educational Psychology* (London), 52, 3 (1982), 341-9.

Past research has often confirmed the poor reliability and validity of oral examinations, but little is known about the factors influencing the evaluation of orals. Two recent studies of the secondary-school leaving examination in Germany investigated the influence of information about past achievement, sequential effects and speed of speaking. Using an analysis of variance design, examination sessions in literature and geography were recorded on videotape and evaluated by 156 teachers of German literature from 39 schools and by 81 geography teachers from 24 schools. The results confirmed a remarkable between-rater variance. The differences were significantly influenced by the speed of speaking (about 17 per cent), background information about past achievement, and the contrast of performance related to the sequence of presentation. Examinations which are so open to uncontrolled influences have a low inter-rater reliability and therefore low reliability and validity. Further research is needed to analyse different sources of differences among various raters. Raters should be trained according to these findings in objective rating techniques.

94. JASMIN-DEMERS, Louise

Des tests pragmatiques et multi-fonctionnels pour une évaluation objective de la compétence à communiquer. [Pragmatic and multi-functional tests for an objective evaluation of communicative competence.] *Bulletin of the CAAL* (Montreal), 4, 2 (1982) 115-24.

Addressing the basic question of the possibility of authenticity in language testing, the article describes two out of seven new tests of French as a second language, derived from Canadian research into the elements that compose authentic oral communication. The first is based on a silent cartoon film called 'Balablok' which the candidate has to explain to a 'naïve' speaker of French. The second (Rallye test) is an interaction game between the candidate and a twinned native speaker.

Numerical results are set out and analysed. Because of the small test population results must be accepted with caution, but it appears that both tests, though very different, intercorrelate and discriminate well and give a reliable measure of communicative competence. While the Balablok interview test, stressing the discourse component of competence, appears the more reliable of the two, the Rallye test seems more authentic, because of the multiplicity of tasks to be performed, and explores more deeply the illocutionary component. One hypothesis confirmed is that reaction time is affected by many more factors than mere knowledge of the language. Each test analyses linguistic competence, which varies as

ABSTRACTS

a function of the discourse or illocutionary act of which observations are taken.

95. JOHNS LEWIS, Catherine

Testing communicative competence. *MALS Journal* (Birmingham), 1981, 19-61.

This paper explores progress made towards testing communicative competence, beginning with a brief overview of types of test. It compares discrete point and integrative tests; explores those aspects of communication that test constructors need to be aware of, if they are concerned to test communicative competence; offers a few critical comments on test items which appear to test communicative competence, but do not do so; and suggests a rationale for assembling tests, when the primary focus is communicative competence.

96. JOHNSTON, Peter

Prior knowledge and reading comprehension test bias. *Reading Research Quarterly* (Newark, Del), 19, 2 (1984), 219-39.

This paper addresses the problem of the effects of prior knowledge, especially those relating to bias, in tests of reading comprehension. Quantitative and qualitative effects of prior knowledge on reading comprehension were demonstrated through an examination of performance on different question types. The availability of the text during question answering was also found to influence performance on certain question types. Peripheral textual items were most sensitive to such influence, central items and scriptal items were least sensitive. Performance on central questions actually improved when readers could not refer back to the text. The biasing effects of prior knowledge were demonstrated both within subjects and between subpopulations (rural and urban). Bias was shown to operate at the level of the individual and therefore should be removed at that level, not at the population level. This was achieved by using a content-specific vocabulary test to estimate prior knowledge, and incidentally resulted in a decrease in the bias due to intelligence. A conventional approach to bias removal (collapsing across several text content areas) also removed the bias due to prior knowledge, but resulted in greater bias relating to intelligence. Results are interpreted to suggest modifications of current reading comprehension tests and methods of dealing with bias.

97. KAISER, Christian

Leistungskontrolle und Fehlerbewertung an Hauptschulen. [Assessment of language performance and error at 'Hauptschulen'.] *Englisch* (Berlin, FRG), 17, 3 (1982), 81-7.

A survey was carried out on English teaching in three Berlin Hauptschulen (non-academic secondary schools) to find out how language performance and errors were assessed. The survey concentrated on defining the areas of language being covered, finding out what types of tests were being used, and ascertaining the distribution of language areas and skills being tested.

The results showed a general imbalance between the approaches of the schools in question, and some evidence of subjective marking. They also raised considerable doubt as to whether the aim of Hauptschule English was being achieved, namely, to promote the pupils' communicative ability in everyday situations.

98. KEITGES, David J

Language proficiency interview testing: an overview. *JALT Journal* (Takamatsu-shi, Japan), 4 (1982), 17-45.

The paper discusses the usefulness and limitations of the direct testing of speaking proficiency in an interview setting, with one or more candidates and one or more examiners. The procedure developed by the US Foreign Service Institute (FSI) is reviewed: it is successful not only because of its established criteria and rating system but also because of the informed subjectivity exercised by its examiners. It cannot, however, be freely transferred to dissimilar testing situations, such as secondary-school and college classrooms. Modifications have therefore been made, such as the use of one instead of two examiners, shorter interviews, testing groups of three to five instead of individuals, and simplifying checklist scales. The interview can also be used for other purposes, such as needs assessment, ability grouping, peer grading, and for establishing and maintaining course goals and standards of teacher training [guidelines are given for the design administration and rating of interviews].

99. KLEIN-BRALEY, Christine

Die Übersetzung als Testverfahren in der Staatsprüfung für Lehramtskandidaten. [Translation as a test-form in the state examination for prospective teachers.] *Neusprachliche Mitteilungen* (Berlin, FRG), 35, 2 (1982), 94-7.

Translation is still the most widely used means of testing the language abilities of finals students aiming to teach. The author

questions and discusses whether a translation test meets the accepted criteria for a good objective test, whether it can really be a valid test of language ability and one appropriate for a future teacher, and whether one's mastering of a language automatically means that one is a good translator. The level of difficulty of texts used for finals translations seems inadequately defined and varies from semester to semester and university to university. Also, it is not clear how freely one should translate and individual examiners have different views on the matter. Some penetrating questions are posed as to exactly how certain mistakes should be assessed.

A comparative table is given showing the results of 26 students in two translation tests taken four weeks apart and marked by the same native-speaker team according to a firmly established marking scheme. The results demonstrate how texts ostensibly parallel in terms of difficulty do not produce the consistent improvement which might be expected. Other methods of testing finals students' language abilities should be investigated.

100. KLEIN-BRALEY, Christine, and Ulrich Raatz

A survey of research on the C-test. *Language Testing* (London), 1, 2 (1984), 134-46.

The use of cloze tests is beset with problems. The C-test represents an attempt to develop a measure of general language competence which will avoid these problems. A large number of studies, involving children and adults learning a variety of languages, point to its being a reliable and valid measure of overall language ability. Construct validation has now begun. The principal usefulness of the C-Test is seen to be in selection and placement procedures. It is essential that the test should not be used to make significant decisions without prior statistical evaluation.

101. KLEIN-BRALEY, Christine

A cloze-up on the C-test: a study in the construct validation of authentic tests. *Language Teaching* (London), 2, 1, (1985), 76-104.

Cloze tests and C-tests are both tests of reduced redundancy based on the theory of general language proficiency. This paper presents the theory and shows first why cloze tests are unsatisfactory operationalisations of the theory and the ways in which C-tests are technically superior. It then reports on the various investigations which have been performed in the construct validation of C-tests and discusses their relevance to the original theory. Four hypotheses are set up relating to linearity, parallelism, prediction of difficulty and processing strategies. The results obtained with the C-tests support these hypotheses.

102. KOSTER, C J

MO-studenten kunnen niet vertalen. [Secondary school students cannot translate.] *Levende Talen* (The Hague), 363 (1981), 578-90.

Very poor results are obtained by secondary-school students in translation tests both from English to Dutch and from Dutch to English. These students perform notably worse than students in universities on translation tasks, even though on other tasks both groups are broadly comparable. Explanations often put forward to account for this discrepancy are considered: (1) the school examiners have pitched the pass mark at the wrong level; (2) the students get too little practice in translation; (3) secondary-school students cannot translate; (4) the teaching of translation is inadequate; (5) it is impossible to learn how to translate; (6) translation is totally different from other language skills; (7) the examiners are careless in their work; (8) the marking schemes used are unsatisfactory. Evidence is presented which leads to the rejection of all these explanations except the first and last, and it is concluded that there is something basically wrong with the examination itself.

103. KRZANOWSKI, Wojtek J, and J Anthony Woods

Statistical aspects of reliability in language testing. *Language Testing* (London), 1, (1984), 1-20.

Reliability is a concept central to testing. Whatever measure of reliability is used, its value will always have to be estimated from a sample and will therefore be subject to sampling variation. The magnitude of sampling error involved in measuring reliability from a sample should be indicated by quoting the standard error of a reliability estimate or by giving a confidence interval for its true value. This paper considers several simple ANOVA models which can be used to define and estimate reliability coefficients. It discusses the statistical properties of commonly used measures of reliability and explains how to test hypotheses about their true values.

104. LANTOLF, James P, and William Frawley

Oral-proficiency testing: a critical analysis. *Modern Language Journal* (Madison, Wis), 69, 4 (1985), 337-45.

The authors question the validity of the assessment guidelines produced by the American Council on the Teaching of Foreign Languages and the Educational Testing Service, particularly in regard to criteria for oral-proficiency testing (OPT). They have not been subjected to sufficient professional scepticism; they are based

on untested assumptions about such factors as the number of hours required to produce a proficient L2 speaker, and the correlation on untested assumptions about such factors as the number of hours required to produce a proficient L2 speaker, and the correlation between topic interest and spoken accuracy. Moreover, there is a fundamental logical circularity at the core of the guidelines, wherein necessary facts are established a priori and then 'proven'.

Learners are gauged in terms of a closed, self-defining system that models rather than mirrors authentic communicative interaction. Despite the criterion-referencing claimed by the guidelines, the rationale really implies norm-referencing. The 'sacred' concept of the educated native speaker, by which L2 learners are evaluated, is artificial, normative and statistically derived. Given the supposed empirical paucity of the guidelines, a call is made for a delay in any decision to implement them in the production of actual teaching programmes.

105. LAPKIN, Sharon

An analysis of French verb errors made by second-language learners in a bilingual programme. *Studies in Second Language Acquisition* (Bloomington, Ind), 2, 2 (1979), 65-84.

Data were gathered using a French cloze test in the context of a longitudinal evaluation of a number of French immersion programmes in Ontario conducted by the Bilingual Education Project of the Ontario Institute for Studies in Education. It was clear from an early stage in the longitudinal study of early French immersion that immersion pupils were attaining a level of second language proficiency far superior to that of the comparison pupils taking core French, and that a more appropriate reference group for measuring French language proficiency would consist of native French-speaking pupils of the same age and grade level. The focus is now on the degree to which immersion students can attain native-like ability in the second language. [Method; results.]

The performance of all groups in the verb category appeared to reflect their performance on the test as a whole: the immersion and Franco-Ontarian groups made similar percentages of incorrect responses whereas the unilingual French group made considerably fewer errors. The relative frequency of erroneous responses as they are distributed between item errors and part of speech errors is similar, although relatively more non-responses are recorded for the IMM group. The hierarchy of difficulty for each verb is similar for all three groups. Where recurrent errors are noted, the IMM and CI groups tend to make the same errors. Moreover, recurrent errors

made by the unilingual francophone group are often among those made by the two bilingual groups. This is consistent with findings for other parts of speech and with the suggestion that cross-lingual interference is not the sole cause of errors among the immersion students. For many of the items examined, a substantial number of recurrent errors can be explained by an overreliance on the immediate phrasal environment which triggers a response that may be both lexically and morphologically appropriate in that immediate context, but not in the larger context of the sentence or the story as a whole. This is a more pronounced tendency in the two bilingual groups; however the fact that it occurs in all three groups suggests that similar strategies are being employed by all students in processing the cloze passage.

106. LARSON, Jerry W, and Harold S Madsen
Computerised adaptive language testing: moving beyond computer-assisted testing. *CALICO Journal* (Provo, Utah), 2, 3 (1985), 32-6.

Computerised Adaptive Language Tests (CALT) have the characteristic of adjusting to the ability level of the person taking the test. Such a facility is needed since item response theory has demonstrated that the most precise measurement of a person's proficiency is one derived from items at or near one's level of ability. Rasch analysis and other Latent Trait procedures attempt to capture an invariable 'yardstick' that measures student ability in relation to item difficulty. This is now made possible by CALT, and one of the software programs for analysis test results on the Rasch model is described, viz Microscale Plus™. It can generate results tables portraying item difficulty and student ability, and maps and diagrams portraying each in relation to the other. It facilitates the linking of items at the same level of difficulty and the creation of calibrated item banks. The computer can administer the test and adjust questions to the candidates' ability level, thus getting a more precise measurement than by conventional means.

107. LEBLANC, Raymond, and Gisèle Painchaud
Self-assessment as a second language placement instrument. *TESOL Quarterly* (Washington, DC), 19, 4 (1985), 673-87.

This article reports a sequence of experiments leading to the use of self-assessment as a placement test. These experiments dealt with such questions as: Do students have the ability to evaluate their own performance? Does the type of instrument used affect that ability? Can students be satisfactorily placed by self-assessment results alone? The conclusion is that at least under the conditions described,

self-assessment must be considered a very valuable tool as a placement instrument.

108. LISKI, Erkki, and Simo Puntanen

A study of the statistical foundations of group conversation tests in spoken English. *Language Learning* (Ann Arbor, Mich), 33, 2 (1983), 225-46.

This paper is intended as one statistical step towards achieving objectivity in group oral testing in English. The test and the marking system adopted were developed by Folland and Robertson (1976). The data under study comprise sample results from tests taken by 698 Finnish university students at the University of Tampere. The testees' performance was evaluated by noting errors under four categories - pronunciation, grammatical structure, lexis, and use - and 'pluses' under use and lexis. The primary aim of this research was to bring out statistical factors which may be useful in developing tests of spoken English. Distributions of errors are shown and studied in relation to variables such as number of utterances spoken by the testee, sex, mark in a matriculation examination, and group factors. The results show that most errors were made in grammatical structure and fewest in use. More talkative students showed better proficiency than less talkative when compared on the basis of the number of errors per ten utterances. Generally considered, the girls' proficiency was better than the boys' but this difference disappeared when the mark in the matriculation examination was taken into account. The mark was positively correlated with language proficiency and the girls' marks were significantly higher than those of the boys.

109. LOW, Graham

Communicative testing as an optimistic activity. *Working Papers in Language and Language Teaching* (Hong Kong), 4 (1981), 1-18.

The term 'use-of-language test' is perhaps preferable to 'communicative test'. It is important to be able to test syntax, where this forms part of a language-related activity that the testee is required to perform, and to see this as a language-use test. A strong case can be made for making not just individual test questions but a whole test or series of questions purposeful. Even if the tester does not adhere to one particular model of competence/performance, he should nevertheless have a clear theoretical stance, even if this is simply to state that a broad view of language competence is being taken. Performance testing is likely to lead to a generally more satisfactory test, it being very difficult to define competence satisfactorily.

The main stages in the production and exploitation of a use-of-language test are, first, establishing what is to be tested. This involves making a profile of what the testee is supposed to do when writing essays, etc., and a statement of what will be tested. Secondly, constructing a good test. Can the directness of a test actually be measured? How should context be manipulated and presented? The ideal test would present the testee with full contextual data; any restriction or selection that the tester cares to introduce for the purposes of testing should be justified explicitly. The tester should realise that 'authenticity' includes the testee's mental processes as well as the content of what is written or spoken. Questions which look like questions, rather than like parts of a language-related activity, are often not perceived as 'authentic' by the testee, who promptly fails to produce the behaviour which the test is aiming to test. The concept of what a test question is and what it should look like needs re-thinking. There is a case for altering the examiner's role to one where he personally engineers problems and communication breakdowns.

A questionnaire designed to accompany the test is helpful in validating it, also interview follow-ups. The question of 'reliability' as a tool of test development. Since the tests involve complex situations and interrelating sets of variables, subtests tend partially to overlap, so total marks do not necessarily represent what the tester imagined they would. The establishment of even remotely adequate performance criteria is the hardest task of all.

110. LOW, Graham D

The direct testing of academic writing in a second language. *System* (Oxford), 10, 3 (1982), 247-57.

This paper attempts to derive, use and partially validate, for one small area of language use, a theoretical approach to the design of language-use tests, called 'explicit direct testing'. A 'direct test' is a test where all significant aspects of a task and the conditions under which it is performed are present. Complete directness is an ideal position; the need for precise assessments of ways in which a test is not direct is the basis of explicit direct testing.

The argument is restricted here to the writing of university tutorial papers or short essays and to considerations of what is called the design criterion, rather than scales of anticipated testee performance, which could be labelled performance criteria. Applying the ideas suggested to the construction of an actual test produced two features which are of general interest to language test designers; the first is the idea of a controlled reduction in directness and the second

that of exploiting a developing context. If the concept of explicit direct testing is to have any real value, however, it must be possible not just to construct tests which have high validity, but to do so without losing all control over the relevant variables. Some results are cited from an evaluation study of the test that was used as the example, which suggest that, on the whole, the approach is feasible for language-use testing.

111. LOWE, Pardee, Jr

The ILA oral interview: origins, applications, pitfalls, and implications. *Unterrichtspraxis* (Philadelphia, PA), 16, 2 (1983), 230-44.

The Oral Interview test devised by the Interagency Language Roundtable (ILA) is described. It was adopted by the US State Department in 1956 but hardly known beyond government agencies until 1978. As an integrative test of proficiency, eliciting functional oral language from the candidate, its wider potential for business, schools and academia is now recognised, and it has been revised with this extended constituency in view by the American Council of the Teaching of Foreign Languages (ACTFL).

The current nine-level rating scale is presented in full and the terminal performance of candidates at each level defined. Six levels from 0 to 5 are recognised and defined in terms of functions, content and accuracy. The top of the scale is educated native speaker performance. Validity is secured by testing job samples of language in simulated real-life situations. Inter-rater and intra-rater reliability is secured by thorough training of examiners; lack of such training lays the system open to misuse and misconception. If this can be avoided, it promises to offer a valid and accurate measure of practical oral proficiency of benefit both to government and academia, not least by virtue of the fact that it provides both with a common scale to which teachers could orient their teaching.

112. MARKHAM, Paul L

The rational deletion cloze and global comprehension in German. *Language Learning* (Ann Arbor, Mich), 35, 3 (1985), 423-30.

The intersentential sensitivity of the rational deletion cloze procedure was examined in this investigation. A total of 84 college-level students of German participated in the study. One group of subjects completed a sequential cloze task, and a second group completed a scrambled cloze task. The results revealed no significant differences between cloze test scores in the exact word or the acceptable word (synonyms allowed) scoring conditions. It is

suggested that the cloze procedure may not yield a valid and reliable assessment of global comprehension in the second-language context.

113. MARS DEN, Robert, and Nick Underhill (International Language Centres). A comparison of some common types of language tests. *Journal of Applied Language Study* (Weybridge, Surrey), 1, 1 (1982), 29-39.

The aim of the study was to compare the reliability and validity of five common types of language tests, and to evaluate alternative marking schemes for two of them. The five tests administered were: (1) A 100 question multiple-choice test, of which most questions were structural but with separate aural comprehension and vocabulary sections. (2) A conventional written cloze test with every fifth word deleted, giving a total of 50 blanks. (3) A 100 word taped dictation based on University of California material. (4) A short essay with a given title: An eyewitness Report of a Crime. The students had to write 6-8 sentences, as if for the police. This particular subject was chosen as being a fairly realistic written task. (5) A teacher's assessment: each teacher of which there were three to four per course, awarded each student a mark out of 50 on the basis of the student's overall proficiency in English. The tests were marked by six trained and experienced EFL teachers. The students were attending four-week holiday courses, were 12-18 years old, and came from Italy, Sweden, Germany and Finland.

The multiple choice test, cloze test and dictation all had high correlations with each other and with a battery. The teachers' assessments did not correlate quite so well and the essay had low correlations with individual tests and with the battery. Alternative marking schemes were tried for the cloze and essay; with the cloze, exact word marking gave only a slight gain in objectivity but reduced marking time considerably. It was though justifiable only for large numbers. With the essays, it was better if a batch of essays was marked by a single marker. Two or three markers should mark each script if possible. There was no great difference in reliability or validity between the impressionistic and analytic marking systems. For up to 50 or so papers the best method was sorting papers into five piles (from diabolical to very good), then ranking the papers in each pile by direct comparison.

In terms of the overall aim of comparing different test types, the single most important result of the re-marking was that the correlations using the new essay scores were as high as those of any other test.

114. MASON, Victor W

Using multiple-choice tests to promote homogeneity of class ability levels in large EGP and ESP programs. *System* (Oxford), 12, 3, (1984), 263-71.

The paper addresses the problem of excessive heterogeneity of ability levels in the courses of large English-language (EL) programmes as a serious concern to classroom teachers and programme administrators alike. The effectiveness of Kuwait University Language Centre's English Placement Tests in promoting homogeneity within courses of both EGP and ESP programmes is discussed, as is use of the English Norming Tests to diagnose the degree of end-of-semester homogeneity within the same courses.

The question of the appropriateness of using 100 per cent multiple-choice tests of grammar, vocabulary and reading comprehension for placement and diagnostic purposes is given special attention, particularly in reference to ESP programmes following the 'communicative' approach. The study concludes that well-written, comprehensive objective tests and direct-proficiency performance tests are complementary rather than incompatible in assessing student achievement and programme effectiveness, and that such objective tests are indispensable both in helping to form homogeneous classes and in acting as a 'check' on the adequacy of any EL programme in providing the basic linguistic skills essential for the development of true communicative competence in English.

115. MASSAD, Carolyn Emrick

International proficiency tests developed for IEA International Study of Achievement in English as a Foreign Language. *ELT Documents* (London), 108 (1980), 24-39.

The International Association for the Evaluation of Educational Achievement initiated in 1965 a cross-national study of achievement in six subjects, including English as a foreign language. An EFL committee was set up to design and develop appropriate achievement and other measures, and tests were administered to nearly 30,000 students, beginning in 1971. The article discusses the considerations that went into the design, construction, administration and scoring of the English proficiency tests used in the IEA study. The analyses contained in the full report of the study, *The Teaching of English as a Foreign Language in Ten Countries*, show that the tests were completely adequate for the purposes for which they were intended and valid for measuring English proficiency in all countries that participated in the study.

116. MITCHELL, Rosamund

The use of role-play tasks in assessing FL communicative performance. *British Journal of Language Teaching*, 23, (1985), 169-72.

Communicative competence includes at least four main elements: grammatical, sociolinguistic, discourse and strategic competence. To be 'certified' as communicatively competent in any foreign language, the learner must actually be required to display this full range of skills, but existing assessment practices focus too narrowly on the learner's linguistic knowledge. Communicative competence can only be inferred indirectly, from actual performance. The problem for the tester is to devise tasks and activities which will elicit an adequate sample of the learner's attempts at communicative performance. Role-play tasks are commonly used for this purpose in graded test schemes, but the criteria of interaction, unpredictability and of real-life performance constraints are frequently not met. More open-ended role-play tasks are needed. Teachers, moreover, need training to conduct such tests. The high-status figure of the teacher, as interlocutor in the test, may have an inhibiting effect on pupil performance.

The author studied an experienced teacher administering open-ended role-play tasks to pupils in their second year of French. She sustained her 'French' roles fairly consistently, but both she and the pupils modelled conversational strategies on previously studied coursebook examples. She varied her responses to different pupils but this partly depended on her prior knowledge of the individual pupil's FL ability and self-confidence. The resulting discourse structure diverged from real-life norms: conversations tended to be constructed out of autonomous question/answer exchanges. The validity of role-play tasks, even open-ended ones, for all-sided assessment of communicative performance, can therefore be questioned, though such tasks undoubtedly have potential for the purpose.

117. MODU, Christopher C, and C Andrew Weeks

The validity of the multiple choice section of the Advanced Placement German language examination. *Unterrichtspraxis* (Philadelphia, Pa), 15, 1 (1982), 116-26.

The Advanced Placement (AP) courses and examinations of the College Board permit college placement and credits to be given for work done at high school. A new AP German programme given in 1979 emphasises 'use of the language for active communication' rather than, as formerly, for the study of literary texts. As a mediator of college credit, the programme depends on the AP

German Language Examination, which must provide an adequate measure of the knowledge and achievement required by second- and third-year college German courses. Validity will mean the examination (a) tests the kinds of knowledge and (b) requires roughly the same level of skill as do the college courses. A study was carried out to check this by comparing performances of AP candidates with those of college students, using a condensed version of the multiple-choice section of the examination. [Test design and administration.] Results showed that the AP candidates, slightly difficult for third-year students and decidedly difficult for second-year students. There were significant statistical relationships between the college students' examination scores and their instructors' ratings of their course performances.

118. MONFILS, Gerard

Etude de la validité du MLAT comme prédicateur du succès aux tests de rendement des fonctionnaires fédéraux canadiens dans l'apprentissage du français, langue seconde. [Study of the validity of MLAT as predictor of the success rate of Canadian civil servants learning French as a second language.] *Medium* (Ottawa), 7, 2 (1982), 21-46.

Aptitude has been defined as the degree to which an individual has mastered the skills requisite for a learning task. According to Carroll, the components of language learning aptitude (which he assumes to be innate) are: Phonetic Coding Ability, Grammatical Sensibility, Memory Ability and Inductive Learning Ability. The battery of tests which constitute the Modern Language Aptitude Test (MLAT) devised and published by Carroll and Sapon in 1959 - Number Learning, Phonetic Script, Spelling Clues, Words in Sentences, Paired Associates - test the first three components. Other researchers have questioned whether language learning aptitude is innate, or have stressed the greater importance of learners' motivation. In the United States, however, MLAT has generally proved to be a reliable predictor of success in foreign-language learning.

In 1977, MLAT's validity was tested in Canada in a different kind of language-learning situation where social and psychological factors outweighed innate ability; MLAT was shown not to be a reliable predictor of success in the case of the group being studied, Canadian civil servants needing competence in French for their work. MLAT did, however, prove reliable in predicting the length of time students took to learn and in distinguishing fast and slow learners.

119. MORROW, Keith

The evaluation of tests of communicative performance. *Prospect* (Adelaide, Australia), 1, 2 (1985), 32-44.

This article first sets out the rationale and specifications underlying the Royal Society of Arts 'Communicative use of English as a Foreign Language' examinations, and then considers criteria for evaluating communicative tests in general. The competence descriptors for the three RSA levels are discussed in terms of such factors as accuracy, appropriacy, range and flexibility; the author emphasises the authentic nature of the tests.

The traditional tests criteria of validity and reliability should be expressed practically; the author coins the terms 'washback' and 'operational' validity. The former refers to the actual effect of tests on particular teaching situations, while the latter concentrates on the internal characteristics of the tests themselves. Operational validity is related to predictive validity.

The scientific and indirect measures which typify classical test evaluation are irrelevant when applied to communicative examinations. A call is made for a distinction between orthodox language testing and that which examines, in a commonsense, rough and ready way, language use by individuals in the real world.

120. MOTHE, Jean-Claude

Evaluer les compétences de communication en milieu scolaire. [Evaluating communicative competence in a school environment.] *Français dans le Monde* (Paris), 165 (1981), 63-72.

Traditional test segmented language either into skills (e.g. reading, speaking) or language categories (e.g. vocabulary, grammar). In life such separation is rare, and tests should seek to establish how far the learner is able to use the foreign language in real communicative situations. Constructing communicative tests for pupils at school is not easy, since their future language needs are not known, but it is not impossible. Comprehension tests are easier to produce, but even communicative tests of production can be devised by using, where necessary, instructions in the mother tongue, and thereby reproducing the specificity of context and intention found in real life. [Numerous examples from *Méthode Orange 2*.]

121. MURRAY, DE

Diagnostic testing of the intelligibility of Thai and Japanese speakers of English as a foreign language. *Working Papers of the*

Speech and Language Research Centre (North Ryde, NSW), 3, 4 (1982), 1-21.

The aim of this study was to explore testing devices which could be used to diagnose objectively which aspects of a learner's segmental phonology cause lack of intelligibility. A brief preliminary study of suprasegmental factors showed that separate diagnostic tests are necessary to evaluate student performance at both the segmental and suprasegmental levels, and that there is a need for objective testing (teachers are not necessarily good judges of oral performance). Two secondary studies investigated the nature of the errors of Thai and Japanese speakers' performances which caused them to be unintelligible. For Thai speakers, it was found that, word-initially, fricatives and affricates, both voiced and voiceless, contribute most to lack of intelligibility, while word-finally, bi-labial and velar plosives are least intelligible. For the Japanese, word initially, the fricatives and affricates were also least intelligible but word-finally the picture is not so clear as there is considerable variation between the results from two versions of the test, due to differences in auditory discrimination. Both groups were found to be least intelligible in their articulation of word-final consonants.

The results from this study were compared with contrastive studies of both Thai and Japanese. This clearly showed that language transfer does not provide a complete explanation of contribution to lack of intelligibility. Many misarticulations predicted from these contrastive studies were in fact intelligible to the judges. For example, /l/ and /r/ do not contribute strongly to lack of intelligibility word-initially for Japanese while Thai articulation of /b/ word-finally is less intelligible than of /d/, but a contrastive study does not indicate this.

122. NIENHUIS, L J A

Schrijfvaardigheidstoetsen voor hogere niveaus. [Testing writing proficiency at university level.] *Toegepaste Taalwetenschap in Artikelen* (Amsterdam), 20, 3 (1984), 80-90.

In this article, the quality of two tests of writing proficiency at university level are discussed. The two tests represent two realistic tasks at this level: summary, and short, informative text. The author investigated empirically the practicability of the tests, and the reliability of rating procedures.

Different groups of 30 university students wrote summaries of two texts, and short, informative texts on three subjects. Each

summary or text was rated by two raters, who worked independently; they rated, separately with an interval of some days, (a) correctness of language, and (b) content and organisation. In order to improve the reliability of rating, correctness of language was rated per paragraph: raters divided each summary in three, each informative text in four paragraphs of approximately equal length, and scored each paragraph separately; while rating the content and organisation, raters had to discern two, or four different aspects, among them the completeness of the content, and the organisation (of paragraphs and) of text.

Raters did not have to analyse texts thoroughly: for each rating they read a text once or twice. In this way, a summary can be rated in $2 + 3 = 5$ minutes, an informative text in $3 + 5 = 8$ minutes. Correlations between ratings of correctness of language and of content and organisation of the same rater were not very high; they were lowest for the informative text; this suggests the well-foundedness of the separate rating of these two aspects. Rater-reliability was higher for correctness of language than for content and organisation. To ensure satisfactory reliability, two ratings are necessary, but even then two raters disagree in about 30 per cent of the cases: these should be reconsidered. It is possible to reduce rating time: the findings show that if rater A only rates correctness of language, and if rater B rates the same texts for content and organisation, they will agree in about 60 per cent of the cases. Each of them should then rate the other aspect for the 40 per cent of students whose work is to be reconsidered. The time left can be spent reconsidering the most difficult cases not yet agreed on. This reduction of rating time is of particular interest if students are asked to write, for the sake of reliability, two (or more) texts.

123. NIZEGORODCEW, Anna

The role of aptitude in foreign-language learning in secondary grammar-school pupils. *Glottodidactica* (Poznan, Poland), 13 (1980), 37-48.

Research is described on working out and applying foreign language aptitude tests for prognostic (testing) and diagnostic (ability level) purposes. Aptitude is defined as the sum of relatively constant, individual features affecting the attainment of a particular level of skill in performing an activity or learning. The research was carried out in two grammar schools teaching English as a foreign language: one form was using an intensive course (I), one an ordinary course (O), but the same teacher taught both forms using the same methods, and pupils were of equivalent general

intelligence. Eleven test techniques were tried of which seven were satisfactory. Results of these tests were correlated with school marks in English over six previous terms.

A substantial difference was found between the (I) and (O) forms. Aptitude correlated with foreign-language marks only in the (I) form. The tests found most predictive of success in the (I) form were: language analysis, sentence closure, sound discrimination, associative memory, and visual sequential memory test. Negative prediction may have a higher value, since a low level of aptitude may impede success despite high motivation. Diagnosis of aptitude is more useful than prediction of success, though the latter may be necessary when admitting pupils to intensive courses. The results of aptitude testing confirmed the influence of two out of the four factors of foreign-language aptitude identified by Carroll, inductive ability and associative memory. The influence of phonetic coding ability was found to be insignificant. The ability called by Carroll 'grammatical sensitivity' is so highly affected by the subject's grammatical knowledge that it can hardly be treated as an aptitude factor. The language analysis test also measures grammatical sensitivity, i.e. the ability to discover grammatical rules and to use them upon new lexical material. [Examples of aptitude tests.]

124. OLSHTAIN, Elite, and Shoshana Blum-Kulka
Crosscultural pragmatics and the testing of communicative competence. *Language Testing* (London), 2, 1 (1985), 16-30.

Since inducing communicative competence is the aim of many language programmes, and tests for entry to such programmes should focus on the intended outcomes, means should be found for testing across cultures on the basis, not of grammar and lexis, but of speech acts and how these are differentially realised from one language to another. This is the field of crosscultural pragmatics.

Various data-collection techniques are reviewed in terms of their relative authenticity and their potential validity for studying intracultural, crosscultural and individual variability in speech-act performance: recording authentic exchanges, role play, open-ended questionnaires and discourse completion tests. The above are speaker-performance tests. However, there are also tests of acceptability from the hearer's point of view, e.g. the authors' judgement test, whereby various ways of rendering requests and apologies are measured on a 3-point scale of appropriacy.

Devising tests of communicative competence requires the establishment of norms within a range of variability in terms both of

production (performance) and reception (acceptability). It also requires a decision on the place of grammar and functions in such measurement.

125. OSKARSSON, Mats

L'autoévaluation dans l'apprentissage des langues par les adultes: quelques resultats de recherche. [Self-assessment of language learning by adults: some research findings.] *Etudes de Linguistique Appliquée* (Paris), 41 (1981), 102-15.

Two groups of adult learners in Sweden and one in West Germany were asked to assess their own competence in speaking and understanding the spoken language. The results corresponded closely with the judgements of their teachers. Differences between teachers seemed to indicate that external assessment could not be considered totally reliable. Students generally tended to under- rather than over-estimate their ability - this was particularly the case with a group of women immigrant workers.

When the students' self-assessments of reading and writing skills were compared with the results of a formal written test these also corresponded closely, confirming that adult learners are well able to judge their own progress even on the basis of the somewhat subjective criteria used.

126. PALMER, Adrian S, and Lyle F Bachman

Basic concerns in test validation. *ELT Documents* (London), 111 (1981), 135-51.

The basic types of test validity (the extent to which a test measures what it is supposed to measure) are briefly reviewed: face validity (the appearance of validity); content validity (the extent to which the selection of tasks is representative of the larger set (universe) of tasks of which the test is assumed to be a sample); criterion-referenced validity (the extent to which a test predicts something considered important), and construct validity (the relationship between a test and the psychological abilities it measures). A procedure called 'multitrait-multimethod convergent-discriminant construct validation' requires gathering data to assess convergent validity (how well test scores agree) and discriminant validity (how much test scores differ).

Types of construct validation studies are described: principal-component analytic studies, and multitrait-multimethod correlational studies of language tests, together with a study which affirmed the hypothesis that two language-use skills, speaking and reading, which differ in both direction and channel, are

psychologically distinct and can be measured independently. Confirmation factor analysis was used to statistically evaluate the fit of causal models to the body of data. This makes it possible to identify and quantify the effects of trait and method on test scores. The findings support Oller's divisible language competence model.

127. PERKINS, Kyle

On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly* (Washington, DC), 17, 4 (1983), 651-71.

In assessing the writing ability of foreign learners of English, two types of test are current and feasible. One (direct) requires the candidate to produce some sequential writing, e.g. a composition, and therefore has construct and face validity because the task is what the test is measuring. The other (indirect) measures the candidate performance of recognition, ordering, reformulation and other tasks alleged to be the components of writing, which permits objective scoring but never requires the candidate actually to write.

The first is therefore a more convincing type of test provided methods of scoring can be made reliable. Three methods are considered: holistic, analytic, and primary-trait. Holistic scoring has the highest construct validity. Threats to its reliability because of differences in judgement between raters can be offset by: training, a behaviour-specific rating schedule, careful monitoring, focusing of the raters, and a variety of other techniques. Analytical methods focus on components of the writing to the detriment of factors relating to the text as a whole, so the gain in reliability is not worth the time, and the loss of construct validity. Similarly, primary-trait scoring relies on the identification of factors from an agreed model of discourse. This is not easy to achieve and the time and effort involved is not worth the gain. Objective tests and standardised tests also fail because they concentrate too much on surface features unrelated to meaning and communication. No scheme is perfect and all suffer from lack of definition of what constitutes good writing.

128. PERKINS, Kyle, and Barbara Jones

Measuring passage contribution in ESL reading comprehension. *TESOL Quarterly* (Washington, DC), 19, 1 (1985), 137-53.

Reading comprehension items are valid to the extent that they measure what subjects have understood of the stimulus material. This article reports an empirical analysis of two administrations of two reading tests: the first time, without the reading passages, and the second time, with the passages. Data from the two administrations were used to calculate the passage dependency of

each test, that is, the extent to which questions can be answered without reading the texts upon which the questions are based. The two tests in this research, Davis Reading Test (Davis and Davis, 1956) and Cooperative English Tests (Educational Testing Service, 1960), exhibited little passage dependency. The stability of item types across the two presentation conditions is discussed, and a hierarchy of item-type difficulty is established using latent trait measurement logits of difficulty. The results of this study suggest that classroom teachers should examine commercially available tests carefully for passage dependency. Furthermore, in constructing reading tests, teachers should avoid writing items that test general knowledge. Instead, teachers should strive to write items that test memory organisation and that reveal whether or not inferences have been drawn.

129. PORTAL, Matthew

The APU testing programme in foreign languages. *Modern Languages* (London), 65, 2 (1984), 95-103.

The article describes the work of the Assessment of Performance Unit of the Department of Education and Science and the tests devised by the National Foundation for Educational Research to monitor the achievement of British students of French, German and Spanish at age 13. In the 1983 survey, listening, reading, writing and speaking were all tested using five different tests in each mode. The tests have to follow communicative testing practice but be compatible with clear-cut scoring and evaluation procedures. The tests are described in detail and they are justified by the results of the first main survey. They successfully combine communicative techniques with non-punitive marking applicable across the 13-year-old ability range, when over 80 per cent of pupils are still studying foreign languages.

(See also section 4.2)

130. PORTER, John H

A cross-sectional study of morpheme acquisition in first-language learners. *Language Learning* (Ann Arbor, Mich), 27, 1 (1977) [publ. 1979], 47-61.

Speech samples were elicited by means of the Bilingual Syntax Measure from 11 children aged 27-48 months covering a wide span of linguistic development. The presence or absence of 11 functors was scored in obligatory occasions and an acquisition sequence was determined using three different methods of speech analysis. Although all three methods yielded a strong intercorrelation among

themselves, no correlation was obtained when these three methods were compared with two previously determined orders of acquisition based on first- and second-language research. It was concluded that the previous order of morpheme acquisition obtained through research on L2 learners was probably an artefact of the Bilingual Syntax Measure testing situation. It was suggested that the sequence of morpheme acquisition as determined by the Bilingual Syntax Measure be subjected to a more rigorous validation with the sequence observed in spontaneous production of first- and second-language learners.

131. POWERS, Donald E

Selecting samples for testing the hypothesis of divisible versus unitary competence in language proficiency. *Language Learning* (Ann Arbor, Mich), 32, 2 (1982), 331-5.

Studies have been designed to test whether language proficiency is a single global trait or whether it is divisible into several distinct competencies. Recent studies are reviewed and compared with a recent factor analytic study of the Test of English as a Foreign Language (TOEFL) in several language groups. The comparison suggests that the characteristics of the sample, especially its proficiency in English, can lead to results that tend to support either a unitary trait or several discrete language competencies.

132. PRICE, Eurwen

Monitoring attainment in the Welsh language in schools. *Journal of Multilingual and Multicultural Development* (Clevedon, Avon), 1, 1 (1980), 41-7.

The work of developing instruments to assess Welsh language skills is set in the context of current concern with assessment of performance in key areas of the curriculum in the schools of England and Wales. The particular context of the Welsh teaching situation is outlined and a distinction made between the priorities of teaching Welsh as a first and second language in primary schools. Plans are outlined to devise instruments to monitor attainment in listening, speaking, reading and writing. An account is given of the preparation of a set of wide-range tests to assess the Welsh reading standards of both first-language and second-language pupils, and the linking of these tests by means of the Rasch model is described.

133. REA, Pauline M

Formative assessment of student performance: the role of self-

appraisal. *Indian Journal of Applied Linguistics* (New Delhi), 7, 1/2 (1981), 66-88.

It is commonly assumed that information of students' language abilities can be elicited only through some type of formal examination. This is not the case. It has been found that students can usefully supplement the information normally derived from conventional sources, and that the technique of student self-appraisal can assist the language teacher in identifying students' strengths and weaknesses, interests and emerging needs. The time has come to take a serious look at the alternative available for purposes of student assessment. The technique of self-appraisal reported in this paper focuses on one approach to the appraisal of student performance in a language training curriculum in which the need for a formative measurement instrument has been established.

134. REISKE, Heinz, and others

Der Grundbaustein (BGS) zum VHS-Zertifikat Englisch. [The First Examination towards the Certificate in English of the German adult education colleges.] *Zielsprache Englisch* (Munich, FRG), 12, 2 (1982), 1-28 (pt.I) and 12, 3 (1982), 9-3 (pt.II).

The issue of a specimen test for the First Examination in English provides an opportunity for twelve contributors to discuss the usefulness of this new examination and its relation to the Certificate. In an introductory article, Reiske recalls that the Certificate, elaborated in the late '60s and early '70s, led to a radical and welcome change in language teaching in adult education colleges. He questions the need for the First Examination as a lower level examination. Its objectives, as shown in its syllabus of notions and topics, are similar to those of the Certificate; 60% of the vocabulary is the same. It would have been better to concentrate on the organisation and methods of teaching. Other critics see the new examination as merely a logical consequence of the Council of Europe projects, regret that the linguistic basis and methodological approach have been insufficiently tested, see in it a possible first step towards a centrally organised examination and would have preferred a start to be made in a minor language instead of English. One critic detects a sexist attitude. The majority of contributors, however, welcome the First Examination as an intermediate goal, likely to have a motivating effect on the many students unable to complete the Certificate course; the similarity of objectives is entirely justified; the actual linguistic load is lighter, yet it has been

shown to be adequate for everyday communication. (See also section 5.1)

135. RESCHKE, Claus

Oral proficiency testing: meeting a variety of needs. *Die Unterrichtspraxis* (Philadelphia, Pa) 17, 2 (1984), 265-78.

The rationale of language proficiency testing is explained against the background of initiatives by the American Council on the Teaching of Foreign Languages (ACTFL) which led to the ACTFL/ETS test. Its use at undergraduate, postgraduate and professional levels in the USA is explained and various types of oral proficiency tests reviewed, illustrated and criticised.

Then follows a description of the oral proficiency test of the American Institute of Musical Studies (AIMS), which trains opera singers to speak German at a summer school in Austria. It is based on the Foreign Service Institute Oral Interview Technique and Evaluation Procedure but two interviewers are used, to increase objectivity. In 1980-84, 1000 students were tested and the test proved to be an accurate and reliable predictor, antedating but eventually to be superseded by the ACTFL/ETS test. Rating scales, levels and descriptors are quoted.

136. RICCIARDI, Joseph, and Gregg Edwards

Investigating occupational communication needs in the perspective of communicative performance testing. *TESL Talk* (Toronto), 13, 1 (1982), 55-74.

A project is described which aimed to specify the language communication needs of a large category of employees of the Public Service of Canada, as an aid to the design and development of tests of communicative proficiency. As there are approximately 54,000 bilingual positions in the Public Service, the main employment categories were investigated one at a time, beginning with the Administrative Support Category, which accounts for 31 per cent of all bilingual positions and includes clerical and secretarial services and machine operators. A preliminary study provided information for the setting up of a list of job tasks which implied some use of language; they involved primarily face-to-face or telephone exchanges, the reading of manuals, notes or correspondence, the writing of notes or correspondence. The list of language activities was then verified and their approximate distribution by types of position was determined, using case-study interviews of 102 employees [interviewers' reporting format is given in an appendix]. Certain 'macro-activities' were found to be characteristic of many of

the job-types sampled; they consisted of groups of micro-activities. This helped to make a link between detailed activities for testing and the positions with which the activities are usually associated. A 'map' of mode interactions and sequences was also constructed, to illustrate typical cycles of language-related activity for various occupations.

137. RIJLAARSDAM, Gert, and Henk Blok

Beoordeling van schriftprodukten door leerlingen: theorie en praktijk. [Assessment of written work by pupils: theory and practice.] *Levende Talen* (The Hague), 365 (1981), 753-66.

There is a case for involving students in the assessment of (mother-tongue) written work because (1) it helps students learn more about writing; (2) it encourages students to be independent of the teacher; (3) students are better able to assess the suitability of a piece written for a student audience; (4) they are often better able to explain mistakes and correct points to their colleagues; (5) the method provides for quicker feedback, (6) allows more written work to be done, (7) frees the teacher from the chore of marking, and (8) allows a piece to be assessed several times.

A number of objections to the scheme are discussed and rejected, and results from some studies relevant to the proposal are briefly summarised. The final section reports an attempt to use student assessments in a classroom: 54 students wrote a 10 page article which was assessed, using a 62 point questionnaire, by the teacher and three other students. Reactions to the task were generally very positive.

138. RIJLAARSDAM, Gert

Het beoordelen van discussievaardigheid in een samenwerkende groep: de kleine praat-denkgroep. [Assessing discussion skills in a collaborating group: thinking and talking in a small group.] *Levende Talen* (The Hague), 386 (1984), 546-54.

This paper reports the development of a formal test of discussion skills in small groups. The value of this type of skill and the need for a formal and reliable test of it are discussed. The test outlined here consists in its final form of a set of 19 criteria against which members of a group are evaluated. These criteria are grouped into four sets dealing with each participant's ability to keep to the point of the discussion, his role in the group, his ability to communicate his point of view, and the quality of his contribution. A slightly different set of criteria are used to evaluate group leaders. The

author claims that the test produces scores which show a surprisingly high level of reliability.

139. RILEY, Philip

The communicative approach and the French CAPES oral examination. *Mélanges Pédagogiques* (Nancy, France), 1981 [publ. 1985], 149-67.

The CAPES is a French postgraduate teaching qualification roughly equivalent to the British PGCE/Dip.Ed. with a complex competitive examination including both oral and written elements; but it is questioned whether this test measures efficiently a candidate's linguistic proficiency or teaching ability. It may be so deep rooted in the French social psyche as to be unsuitable for testing teachers of English.

The linguistic, pedagogic and psychological presuppositions of the CAPES system are critically examined in turn. Teaching is seen as a narcissistic process, modelling a future elite on a present elite, which holds up an Oxbridge accent as the norm and equates academic virtues with teaching abilities, ignoring most of the sociolinguistic and technical advances of the last generation, e.g. the tape recorder. Its whole purpose is to exclude, not to train.

A revised CAPES should not be discipline- or content-based but should introduce students to applied linguistics, educational psychology and philosophy and a range of teaching concepts and techniques, with classroom practice opportunities to improve one's language. The oral examination should then only test the candidate's language-listening comprehension, oral expression and interaction.

140. ROE, P

Evaluation in the context of a communicative approach. *Revue de Phonétique Appliquée* (Paris), 61, 3 (1982), 191-200.

A system of evaluation is proposed which is appropriate to a truly 'communicative' approach to foreign-language teaching. The author's working definition of this approach is one whereby a student is trained to develop his realisation strategies to match his use of text to his communicative purpose most effectively. He needs a genuine purpose identified by himself and an outcome directly related to the purpose. The smallest unit of 'communicative act' is 'negotiation', e.g. letting the cat in at night, buying a packet of cigarettes. The definition can now be amended to one whereby a student is trained to perform specified kinds of negotiations or tasks.

Current classroom practice, however, does not often involve the learner in this way.

Evaluation in the context of a communicative approach measures the learner's ability to perform certain kinds of negotiation or task. It is concerned with specifying the level of complexity at which a learner can attempt a given class of task and the degree of effectiveness with which he can achieve it. The information supplied by the system should be valid and reliable, and readily comprehensible to all users. This information flow should constitute a continuous feedback system enabling the learner to monitor his progress at all times. This implies two dimensions of evaluation: task complexity (the number of levels corresponds to the class of task), and level of competence (a 10-point scale is proposed). [An application of such a system is described.]

141. ROGER, Derek, and others

The construction and validation of a questionnaire for measuring attitudes towards learning foreign languages. *Educational Review* (Birmingham), 33, 3 (1981), 223-30.

A questionnaire was constructed for measuring the attitudes of schoolchildren towards learning French. Factor analysis revealed a primary factor accounting for over 60 per cent of the total variance, and an empirical validation study showed that the attitudes of children who obtained high scores on teachers' ratings of attentiveness and enthusiasm were significantly better than those given low ratings. A subsequent replication using a large sample of pupils provided a clarification of the factor structure, and the results of an independent study of attitudes towards learning German showed that the questionnaire can easily be adapted to measure attitudes towards learning other foreign languages as well as French.

142. ROSSI, Gary J

Foreign language proficiency testing: a selected, annotated guide to standardised instruments, rating scales, and literature. *French Review* (Baltimore, Md), 56, 6 (1983), 840-58.

An introductory guide to existing tests of foreign languages and English as a foreign language, with a supporting bibliography. Part I provides short descriptions of the major standardised language proficiency tests available in USA and Australia for French and English. The content, target constituency, rationale, rating scales, and sources of further information are given in catalogue form. Part

It is an annotated bibliography of 94 publications relating to language testing issues.

143. SAVIGNON, Sandra J

Dictation as a measure of communicative competence in French as a second language. *Language Learning* (Ann Arbor, Mich), 32, 1 (1982), 33-51.

Dictation has been proposed as a relatively simple yet reliable and valid indirect measure of functional language skills and is gaining in popularity as a test of placement and proficiency in L2 programmes of all kinds. Yet the continued reliance on exact word scoring of dictation would seem to limit its use as a measure of communicative competence. This paper reports on the development of other than exact word procedures for scoring dictation tests in French. Three separate passages were dictated to college students of French at various levels of proficiency as well as to smaller groups of native speakers and children who had acquired French in 'natural' or immersion settings. A scoring procedure was developed which evaluated chunks rather than words on three criteria: exact word, phonetic similarity, and conveyance of meaning. The analysis of results suggests that this procedure offers a reliable and efficient alternative to word for word scoring on an exact word basis and that the distinction between the three criteria is potentially useful in diagnosing functional 'language skills in classroom programmes.

144. SCHAEFER, Carl F

The cloze procedure for placement testing. *Glottodidactica* (Poznan, Poland), 15 (1982), 75-82.

Notwithstanding the conflicting claims and counterclaims made for cloze tests, they are a reliable and efficient instrument for placement testing as is shown by experience at the American University, grading international students for an English language programme. Students are usually rated in three ways: a teacher assessment on a 1-7 scale; the McGraw-Hill CELT test; a composition writing task. Some also arrive with TOEFL scores that further indicate their level. Two cloze tests were administered to a total of 238 students and scores were shown to correlate at above 0.80 with scores obtained by the other instruments. Scoring was by the acceptable word method on a total of 75 deletions at six-word intervals. The cloze thus shows it is as valid and reliable as other placement tests, if not more so. Moreover, it is easy to construct, administer and mark. Therefore, for placement purposes, the cloze has shown itself to be a powerful and inexpensive test.

ABSTRACTS

145. SCHEIBNER-HERZIG, G, and others

A study of foreign language achievements and personality variables in 15-year-old pupils. *ITL* (Louvain), **65** (1984), 43-56.

Using standard tests with an EFL test specially developed for the purpose, 112 15-year-old German students of EFL were assessed for performance, intelligence, motivation and personality - and the relations between the various scores examined. The tests are described and the measures tabulated, correlated (Spearman) and subjected to factorial analysis by varimax rotation. Five factors are found to contribute most to achievement: spatial and numerical reasoning independent of education 20%; verbal intelligence dependent on education 18%; a well-balanced personality profile 17%; altruistic, social and artistic interest 13%; and business and administrative interest 13%. Thus the personality and interest factors explain more of the language achievement of the pupils than the reasoning and language measures. Listening comprehension is a more powerful determinant of language success than reading comprehension.

146. SCHLISSINGER, Jacqueline

L'évaluation continue et son rôle dans l'accélération des processus d'apprentissage. [The role of continuous assessment in accelerating learning.] *Français dans le Monde* (Paris), **165** (1981), 24-32.

Questionnaires completed by both trainers and trainees following a six months' course in French-as-a-foreign-language for Saudi Arabian students - to be followed by six months' vocational training - indicated the need for continuous assessment. Appropriate evaluation procedures can contribute to the success of language courses by inculcating a serious sense of purpose, checking at every stage on progress achieved, promoting more active student participation in the learning process and ensuring that courses continue to meet students' needs. All aspects should be assessed, including the social side and the practical organisation of the course.

147. SCHULZ, Renate A

Testing in post-secondary self-paced instructional programmes. *Canadian Modern Language Review* (Toronto), **39**, 2 (1983), 243-56.

Individualised language programmes face some special problems in the area of testing and evaluation, among them the need for pre-instructional student evaluation, for increased time and expertise in the area of testing on the part of the teacher, for special test facilities and test security, and for testing procedures which facilitate internal programme mobility. The article presents a model for testing in individualised programmes, including aptitude, placement,

achievement, and proficiency measurement. The author addresses the incongruity between proclaimed course goals and contents of achievement tests and calls for the measurement of discrete-point knowledge as well as the communicative ability and cultural insights of students. The predominant practice of giving only formative tests in self-paced courses is challenged, since this practice does not take into account individual differences in memory, intellectual synthesising ability, and long-term mastery. To protect the quality and reputation of individualised courses, teachers need to keep careful records of enrolments, attrition, grades, aptitude scores, time spent in learning, credit hours earned, and summative achievement and proficiency scores of students enrolled in those courses.

148. SCHWARZ, Michel P M

Tests de langues vivantes: tests, tests de savoir, tests de savoir-faire. [Foreign language tests: tests, tests of knowledge, tests of 'savoir faire'.] *Français dans le Monde* (Paris), 165 (1981), 48-57.

Foreign-language tests are necessarily imprecise: there is no exact yardstick by which to measure performance, and the final score is often a sum of disparate parts; selection based on test results is thus unjust. Criterion referenced test are justified however, provided their validity is proved. What should be tested is communicative competence, subdivided into correctness, appropriateness and pragmatic performance. In the project described here, an evaluation form calls for a judgement by three examiners whether a candidate has reached a given level under the general and the three specific headings above. In all the tests the context is fully specified and the expected performance described [examples].

149. SCHWIBBE, Gudrun, and others

Zur Erfassbarkeit fremdsprachlicher Kompetenz mittels kontentanalytischer Indikatoren. [Evaluating foreign language competence using content-analytic indicators]. *Die Neueren Sprachen* (Frankfurt am Main), 84, 2 (1985), 174-90.

The study in question discusses the inadequacies of subjective evaluations of language performance in essay writing and suggests instead an objective, content-analytic procedure. This procedure consists of parameters which record word redundancy, word length and word fluency, the abstractness and the adjective-verb ratio as well as the error count. The validity tests prove that this procedure is sensitive to the dichotomy of foreign vs. native language competence as measured by the DELTA language test. The advantages of this

procedure with respect to economy, objectivity and reliability are evident.

150. SCOTT, Roger

Communicative evaluation with particular reference to self-evaluation. *Bulletin CILA* (Neuchâtel), 40 (1984), 42-52.

The language needs of a heterogeneous group of foreign learners of English were assessed and a nine-level scale of proficiency was devised for them. Evaluation of the students' performance of language tasks, i.e. their ability to communicate, has to yield information about their ability to use the language they need in the context they require it for. The context is a communicative one and will contain such characteristic features as information gap, feedback and choice of language; the variables to be controlled include size of text, complexity of text, range, speed, flexibility, accuracy, appropriacy, independence, repetition and hesitation. Communicative evaluation will not comprise discrete-point, norm-referenced tests because such testing does not yield the information required.

The most direct way to find out if someone can do something is to get him/her to do it. Regular communicative assessments should be incorporated in the teaching programme. Some suggested modes of evaluation are: role-plays; simulations; letter-writing; listening and giving an oral account of what has been heard; reading and giving an oral account of what has been read; listening and reading and giving a written account of what has been heard. Ways of combining tasks in one activity are suggested. A self-evaluation grid is presented which provides a simple and manageable profile.

151. SEATON, Ian

Principes de la conception de tests destinés à l'évaluation préalable d'étudiants étrangers venant étudier en Grande-Bretagne. [Principles of test design for an initial evaluation of overseas students in Britain.] *Etudes de Linguistique Appliquée* (Paris), 43 (1981), 67-78.

Language tests are changing from discrete item tests, based on a structural analysis of language, to communicative performance tests, related to the testee's language needs. Objective tests satisfy the criteria of reliability and validity, though the latter is in the end determined by real performance. Communicative tests, however careful the description of needs may be, can never predict language behaviour fully. The two approaches can be combined in a model based on a precise description of criteria of performance, down to its

linguistic levels, and satisfying the criteria of reliability and validity through its ability to predict real performance. Such a model has inspired the new English Language Testing Service, designed by the British Council for overseas students in establishments of higher education in Britain. It is the result of a detailed analysis of performance criteria and language use in the life sciences, medicine, physical sciences, social sciences, technology and the humanities. The battery comprises multiple choice tests and 'authentic' integrated activities, common general tests of written and aural comprehension and modular tests differentiated according to disciplines and concerned with study skills and appropriate written and oral tasks.

152. SHOHAMY, Elana

The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning* (Ann Arbor, Mich), 33, 4 (1983), 527-40.

The paper first discusses the complexity of measuring oral proficiency in communicative situations. The difficulty is due to the large number of variables, linguistic and social, which interact with one another. It then reports on a study which examined the stability of the assessment of oral proficiency on the oral interview test. Students of Hebrew as a foreign language underwent four administrations of different versions of that test. The administrations differed from one another by the occasion, the interviewer, the speech style, and the topic. Results from the analysis indicated that the different speech style and topic significantly affected students' scores on these tests while the occasion and the interviewer did not. The correlational analysis between pairs of tests pointed to low reliability and lack of stability of the tests, especially when two variables (i.e. occasion and tester) interacted. The results call for use of caution when decisions about individuals are made based on administration of communicative tests, for a need to identify sources of error in communicative tests, and for drawing stringent guidelines for the use of such tests.

153. SHOHAMY, Elena

Does the testing method make a difference? The case of reading comprehension. *Language Testing* (London), 1, 2 (1984), 147-70.

The paper reports results of a study which examined the effect that various methods used for testing reading comprehension have on the measurement of that trait. The methods investigated in this

research were multiple-choice and open-ended questions, each presented in L1 and L2, on the same L2 texts. Results indicated that each of the testing methods produced different degrees of difficulty for the test takers, and that each of the variables - method, text and language - had a significant effect on students' scores in reading comprehension. These effects were strongest on low-level students. Recommendations are made with regard to the use of multiple procedures for testing reading comprehension as well as to the need for further research towards finding the most valid method for testing that trait.

154. SHOHAMY, Elana, and Thea Reves

Authentic language tests: where from and where to? *Language Testing* (London), 2, 1 (1985), 48-59.

The authors survey test developments, particularly in assessing oral performance, from the statistical emphases of the psychometric-structuralist era to the present-day interest in authenticity. Indirect testing, wherein linguistic competence is sampled artificially, is contrasted with direct testing procedures, which are aimed at duplicating as closely as possible real-life communicative situations. Indirect tests have the disadvantage that they yield no diagnostic information; testees may aim at coping with the peculiarities of the test, not at practising the linguistic traits that the examiner wishes to uncover. Direct tests seem to be more trustworthy and humane but they can produce unstable, unreliable results and also display an artificiality of their own. Because real-life interaction necessarily differs from the communicative patterns found even in direct assessment, authentic tests are themselves capable of eliciting a 'test language', rather than samples of genuine performance. Such tests therefore need to be interpreted cautiously by users. The ideal testing situation would entail unobtrusive observation of everyday communication, though this might well be impossible in practice.

155. SIMMONDS, Paul

A survey of English language examinations. *ELT Journal* (London), 39, 1 (1985), 33-42.

The examinations of 14 bodies (2 USA and 12 British) aiming to test *general* English are impressionistically compared. They are shown to examine differentially the various skills, and to serve different ends and constituencies of candidates. Some 18 examinations cover listening, speaking and writing, and 17 of these test reading also. Grids show the spread of coverage and the test

elements are evaluated against each other and against current testing and teaching notions. They are shown to fall into two groups: 'traditional' and 'communicative'. While this does not imply the superiority of one over the other, the tests that place most emphasis on meaning and the processing of information in a meaningful context are tentatively considered better than the rest, e.g. RSA/CUEFL and ARELS/PRELIM. The conclusions are based only on stimulus and question material, not on candidate response data.

156. SKEHAN, Peter

Language aptitude - a review. *English Language Research Journal* (Birmingham), 1 (1980), 92-112.

This article reviews the field of language aptitude testing and tries to provide an up-to-date survey of theoretical and empirical developments in the area. An examination of the concept of language aptitude is followed by an account of research findings in the period since the publication of the Modern Language Aptitude Test and the Language Aptitude Battery. This is followed by an examination of current theories of language aptitude. The final section attempts to place language aptitude testing into some perspective in relation to current developments in applied linguistics.

157. SKEHAN, Peter

Issues in the testing of ESP. *MALS Journal* (Birmingham), 8 (1983), 61-85.

The case made by Oller for the unitary competence hypothesis is set in contrast to the subdivision of language into discrete skills, particularly that proposed by Munby and exemplified by Carroll in the ESP sections of the English Language Testing Service (ELTS).

Whereas the strong form of the hypothesis is now discredited, a weaker form survives in the consensus that there is something approaching a general factor in language competence which accounts for some of the variance in language tests. This is certainly preferable to the conceptualisation of language as some 54 separate skills/functions or 238 subskills proposed, somewhat arbitrarily, by Munby without empirical justification. Even supposing they are correct, the problems of sampling all of them evenly in a test are enormous - and they can be still further subdivided, for instance, to allow for differences in individual lecturing styles and the accompanying intonation patterns. Practicality aside, however, the ELTS construct validity is questionable because of both its unsound theoretical base and its disregard of validity statistics.

While the Munby model has ostensibly made 'criterion-referencing' feasible by closely defining target behaviour, its purely

linguistic definition fails to account for the complexity of real language behaviour; so the best we can hope for is 'criterion-related' testing suited to each particular ESP situation.

158. SKEHAN, Peter

Issues in the testing of English for Specific Purposes. *Language Testing* (London), 1, 2 (1984), 202-20.

Comparison is made between tests based on Oller's unitary competence hypothesis (UCH) and ESP tests deriving from the Munby model such as the British Council's English Language Testing Service (ELTS). While the strong form of the UCH has now been abandoned, it has led, because of the empirical research and testing done with it, to a widely accepted weaker version where a general factor is seen to coexist with more specific 'skills'.

By contrast, ESP testing has relied on an intuitively selected range of skills very much open to challenge by empirical research and too arbitrary and complex for most practical purposes. While the search for criterion-referencing is attractive, it is only feasible when very narrowly defined jobs are to be tested. The best language testers can hope for is 'criterion-related' testing, so the requirement now is either for the development of appropriate norm-referenced tests or the systematic use of criterion-related validity.

159. SMITH, Karen L.

The USC proficiency-based program in foreign language instruction: methods and their effects on achievement and motivation. *Unterrichtspraxis* (Philadelphia, Pa), 17, 2 (1984), 240-54.

This describes a test in foreign languages given at the University of Southern California after 200 hours (or 3 semesters) of study, and aimed at concentrating the weight of teaching and learning on reading and oral proficiency, counteracting the influence of the multiple choice structure tests of the past. It has resulted in better achievement in Spanish and lower dropout rates through the semesters. The conversation test is an interview and is evaluated on a ten-point global rating scale, which ranks students of the degree of communication achieved. The FSI/ACTFL/ETS scale was considered and abandoned for various practical reasons. The reading test is a multiple-choice test of general reading comprehension. The internal timetabling required to accommodate this shift in emphasis, and the teaching needed for the reading test, are described. Since the test stresses communicative skills, merely 'teaching the test' will now be beneficial.

160. SPOLSKY, Bernard

The limits of authenticity in language testing. *Language Testing* (London), 2, 1 (1985), 31-40.

Setting authenticity as a criterion raises important pragmatic and ethical questions in language testing. Lack of authenticity in the material or method used in a test weakens the generalisability of results. Any language test is by its very nature inauthentic, abnormal language behaviour, for the task is not to give so much as to display knowledge. With examinees who do not know or who are unwilling to play by the rules of the game, the results of formal tests will not be an accurate and valid account of their knowledge. Only part of this difficulty can be overcome by authentic-seeming tasks. Observation of authentic behaviour (even allowing for the observer's paradox) is another partial solution. Long, patient and sympathetic observation by observers who care to help seems the only full solution.

161. SPOLSKY, Bernard

What does it mean to know how to use a language? An essay on the theoretical basis of language testing. *Language Testing* (London), 2, 2 (1985), 180-91.

Language testing and linguistic theory must both try to define language knowledge and use. There are three main approaches. The Structural claim, which assumes that knowledge can take the form of a grammar or structural description of the language, forms the basis of the discrete point tests. The Functional approach assumes that the nature of language knowledge is best captured by listing the various uses to which it can be put; it is embodied in the communicative competence model, the notional-functional curriculum, and the interest in teaching and testing pragmatics. The General Proficiency claim is based on the notion that individuals vary in possessing measurable amounts of an indivisible body of knowledge. It underlies arguments for a general factor underlying batteries of tests or for the trait measured by certain test methods like the cloze. The theoretical strengths and weaknesses of these approaches and the impossibility of showing that any one is completely correct forces us to consider all three as basic to testing. Anybody who knows a second language must be assumed to have all three kinds of knowledge, so that we can only achieve the full picture of language proficiency if we use many different measuring methods and know what trait is being tapped by each test.

162. SPURLING, Steven, and Donna Ilyin

The impact of learner variables on language test performance. *TESOL Quarterly* (Washington, DC), 19, 2 (1985), 283-301.

Research has shown that learner variables such as age, language background, field of study, and educational background have an impact on non-native speakers' scores on language tests. However, does this impact reflect a natural advantage of one group over another, or is it indicative of test favouritism or test bias? This article looks at how the learner variables of age, sex, language background, high school graduation status, and length of stay in the United States affect performance on six language tests: two cloze tests, a reading test, a structure test, and two listening tests. A multiple regression approach to the problem, posed by analysis of variance, was taken to determine (1) the amount of variance accounted for by each of the learner variables on all tests combined and on each test individually and (2) the interaction of variables with test type. Such an interaction would be evidence of either test favouritism or test bias. It was found that age, language background, and high school graduation status accounted for small but significant amounts of variance on the tests. Furthermore, there was an interaction between the test and learner variables. Therefore, both test advantage and test favouritism/bias were found.

163. STANSFIELD, Charles

Reliability and validity of the secondary level English proficiency test. *System* (Oxford), 12, 1 (1984), 1-12.

The SLEP is a group administered test of English language proficiency which includes two subscores and eight different item types. It is designed to assess a foreign student's readiness for English medium instruction at the secondary level. This article describes the development of the test specifications and the performance of each item type and the total test during an administration of SLEP to students in other countries. Several innovative formats are described and discussed, including multiple-choice cloze and multiple-choice dictation. The paper also reports the findings of a validity study which involved the analysis of test scores and demographic data for US public school students.

164. STANSFIELD, Charles W

A history of dictation in foreign language teaching and testing. *Modern Language Journal* (Madison, Wis), 69, 2 (1985), 121-8.

Dictation is one of the oldest techniques for testing progress in foreign-language learning. It has long been associated with the

traditional or grammar translation method. It was strongly rejected by Gouin who advocated the natural method during the second half of the nineteenth century. It regained popularity at the end of that century when the direct method came into favour: phonetic dictation (phonetic transcription of the spoken language) became a common classroom activity. At the beginning of this century, dictation began to appear in standardised tests of modern languages. A study by Rollo Brown in 1915 of teaching methods in France came out strongly in favour of the effectiveness of dictation in teaching composition, and his findings were influential up until the Second World War. It was also promoted in the reading method which was widespread during the second quarter of this century, and during the 1950s.

During the 1960s, the use of dictation declined sharply because of the development of the audiolingual method, which sprang from the favourable results achieved by army training methods during the Second World War. The new methodology was based on the theories of structuralism and behaviourism, with which dictation did not appear to fit in. It was also criticised from the standpoint of testing for not being a discrete-point test, and was regarded as being too complicated and specialised. Valette (1964), however, found dictation to be a good measure for overall proficiency.

Interest in dictation returned again in the 1970s as reflected in Oller's research into integrative measures of language proficiency. It was found to be a good measure of listening comprehension and overall proficiency, and simple to construct, administer and score. Oller posited that dictation taps the learner's internalised grammar of expectancies at work during the listening process. At present, dictation is widely used in both language teaching and testing.

165. STEVENSON, Douglas K

Language testing and academic accountability: on re-defining the role of language testing in language teaching. *IRAL* (Heidelberg), 19, 1 (1981), 15-30.

There are 'restricted' and 'expanded' views of the role of language testing. In the restricted view, language testing is seen primarily within its classroom context: a test tests the pupil, not the teacher. The expanded view is more aware that the measurement of language is at the core of the scientific approach to the study of language behaviour, and is interested in other approaches, such as essays, oral interviews, sociolinguistic procedures, teacher judgements and student self-evaluation. It realises that language

measurements must get to grips with real-life data. Purpose and use will determine the degree of validity of a test, not its title or type. Those holding the expanded view are usually critical observers of tests and are concerned to protect standards.

Germany is moving away from the traditional view but not yet firmly committed to the modern. The limitations of language testing are directly connected to lack of precision in linguistic (and educational) theory. Language measures used in research must not be exempted from considerations of reliability and validity, or research theories will be built on crumbling foundations (e.g. much research based on the Bilingual Syntax Measure). Careful and complex research is often misleadingly over-simplified on its way to the literature of language teaching. Teachers and trainee teachers need sufficient familiarity with basic measurement principles and statistics to be able to interpret what they read in professional journals. They also need training in the administration and interpretation of large-scale, standardised tests, about which there is much misunderstanding. These tests offer relatively inexpensive, efficient, and readily available common element for comparisons; they can also be used to test teachers' consistency in scoring.

166. STRECKER, Bernd

Verfahren zur Messung freier Sprechfertigkeit in der Fremdsprache. [Methods of testing oral fluency in a foreign language.] *Die Neueren Sprachen* (Frankfurt am Main), 80, 2 (1981), 125-42.

In order to establish an objective way of testing fluency in the spoken language, the 'atomistic' and 'global' methods of assessment are compared. When the atomistic approach is adopted, language is analysed in isolated, mutually independent elements; when the global method is used, these elements are seen as integrated into an organic whole.

Authentic speech achieves its goal (the free, creative communication of ideas) with the help of such things as gesture and mimicry, which cannot be quantified or assessed by the atomistic method, the scope of which is restricted to such elements as intonation, grammar, phonetics and vocabulary and sentence structure used. The global method has the advantage that the various elements which contribute to successful communication are considered in the wider context of the general impression made by the speaker. The best way of measuring oral fluency combines both these methods [such a test is described]. The individual components of linguistic communication should be examined in relation to the effect of the whole speech.

167. SWAIN, Merrill

Teaching and testing communicatively. *TESL Talk* (Toronto), 15, 1/2 (1984), 7-18.

Three principles for communicative testing are stated: start from somewhere; concentrate on content; bias for best. The starting point should be the now familiar fourfold subdivision of communicative competence into grammatical, sociolinguistic, discourse and strategic competences. Content includes both the language to be elicited by the examination tasks and the tasks themselves; each task should require all four competences to be drawn upon. The test should be conducted and scored in such a way that the candidate gives his best possible performance. The scoring criteria cover all four competences in each of the tasks and are a mixture of objective counts and subjective judgements. The operation of these principles is illustrated with the material produced to test Grade 9 French immersion students in Canada and it is stressed that the principles assume a pedagogical function for language testing, communicative teaching and testing being two sides of the same coin. Indeed, this testing material could also be a teaching unit.

168. SWALES, John

Examining examination papers. *English Language Research Journal* (Birmingham), 3 (1982), 9-25.

Instructional verbs acquire specialised meanings in the context of examination papers, e.g. 'calculate' implies that graphical methods are not acceptable. Overseas students need attuning to the hazards of written examinations, as they tend to answer questions which were not set, on the basis of recognising one or two words on the exam paper. They will avoid questions containing words or sentences they do not understand, thus their choice may well be much more limited than that of their British counterparts.

A pilot study of six examination papers in chemistry, made by the author, found seven categories of question-types, the most popular being 'describe', followed by 'define' and 'explain'. Each question has an average of three parts and each part has a single or double instruction ('describe...and illustrate...'), i.e. there are 15-20 part answers to complete. About 25 per cent of these can be answered at the word or phrase level, another 15-20 per cent are definitions; about 50 per cent are the 'describe'/explain' types requiring the capacity to construct short but adequate descriptive/explanatory paragraphs; the remaining 10 per cent or so will involve argumentation and evaluation at a higher cognitive level. The

overseas student thus cannot avoid coming up against 'describe'/'discuss' questions. A profile of exam writing tasks should therefore be established.

169. THOMANECK, J K A, and G J A Burgess

Testing in the language laboratory: an argument for its efficiency. *British Journal of Language Teaching* [formerly *Audio-Visual Language Journal*] (Birmingham), 18, 1 (1980), 35-9.

A preliminary report of the findings of an experiment conducted with a group of second-year students of German at the University of Aberdeen. The experiment was designed to ascertain whether the performances of the same students in the same tests improved or deteriorated when those tests were conducted in the language laboratory as opposed to the classroom. At the same time, a number of possible contributory factors was analysed in order to measure and, if possible, eliminate their influence on the results obtained. The form of tests chosen was that of a dictation and a listening comprehension. A clear indication was obtained that a definite improvement overall resulted from conducting these tests in the language laboratory rather than in the classroom. Further experiments along these lines with other groups of learners are planned.

170. TORBE, Mike, and others

The uses of English. *Times Educational Supplement* (London), 3384 (1 May 1981), 19.

About half a million candidates sit O-level English examinations this summer term (1981). There may be grounds for concern about candidates from ethnic minority groups such as Blacks or Asians, many of whom persistently fail this examination despite passing many other subjects at O- and A-levels. A working party of teachers from Coventry wrote to the Vice Chancellors of all British Universities to ask why a pass in O-level English Language was considered essential by those who required it: what they felt such a pass told them about a candidate's communicative competence; and, if they had dropped the requirement, why?

Thirty replies were received. They varied widely but the general implications were clear. O-level English Language gets some of its status from the widely believed idea that universities require it. But it was clear that the majority do not, and that some of them endorse it without much enthusiasm. The same examination is being used by very different people (employers as well as universities) for totally different purposes. It cannot assess communicative competence

satisfactorily for a wide range of different contexts, only one limited form of literacy in a highly specialised context. The examination should enable candidates to demonstrate their use of English in a range of assignments, so that their abilities in varied uses of talking, writing and reading can be assessed, according to clearly defined public criteria of success and competence, which are widely known by teachers, students, parents and users of examination results.

171. VERHOEVEN, Ludo, and Anne Vermeer

Assessment procedures of oral second language proficiency: spontaneous speech vs. elicitation data. *Inter-language Studies Bulletin* (Utrecht), 8, 2 (1984), 50-68.

This article compares the utility of direct versus indirect testing techniques in assisting Dutch teachers to evaluate individual proficiency differences. The concurrent validity of both techniques is established in this study by means of teacher evaluation of native and non-native Dutch-speaking pupils. The linguistic model at work is a hierarchically built grammatical system that distinguishes ability levels in phonological, lexical, syntactic and text-processing terms.

The authors review previous research in validating direct (i.e. spontaneous) and indirect (elicitation) L2 oral assessment, and then describe a study which analysed the abilities of Turkish and Dutch working-class students. The elicitation test procedures included such tasks as Auditory Discrimination, Sentence Comprehension/Imitation and Text Comprehension. The operation of each task is described, as are the results [tabular data].

It is concluded that spontaneous speech and elicitation measures correlated well, both with each other and with subjective teacher evaluation, but indirect measures related better to such 'external' judgements. A battery of easily administered and interpreted elicitation tests is the best way of identifying and diagnosing those learners experiencing linguistic difficulty.

172. VOLLMER, Helmut J

Why are we interested in 'general language proficiency'? *ELT Documents* (London), 111 (1981), 152-71.

Research at the Max-Planck-Institut für Bildungsforschung in Berlin aims to study evidence for the 'unitary competence' and 'divisible competence' hypothesis - to discover whether all performance in a second/foreign language can be traced back to a single underlying factor, 'general language proficiency' (GLP). Two different lines of research are (1) the divisible competence hypothesis

(the assumption that there are a number of specific competencies underlying language behaviour) - this approach has given rise to discrete-point tests. The term 'overall proficiency', or GLP, is not used; the aim is to find the learner's specific proficiency in terms of the content area defined and the criteria used. There is an implied need for some judgement of a learner's 'transfer ability', i.e. ability to use the language being tested in another situation. (2) The notion of 'overall proficiency' as used by Spolsky is misleading and seems to refer rather to an 'underlying linguistic competence'. Oller and others believe that unitary linguistic competence is the principal factor underlying all language skills. However, much evidence suggests that language as comprehension and language as production are profoundly different, even if there is insufficient evidence to claim that there are basically two distinguishable competences.

GLP may perhaps be the centre of a person's general cognitive apparatus (about which admittedly little is known), but in this case should it not be called 'global'? If foreign-language learning is a dynamic process, testing proficiency means making a cut at a given point in order to form a rough idea of a person's progress. Testing should therefore be carried out repeatedly, preferably by different means.

173. WELLER, Franz-Rudolf

Schülerwettbewerb Fremdsprachen - Eine Förderungsinitiative des Stifterverbands für die Deutsche Wissenschaft. [Foreign languages competition in schools: an initiative of the Association of Sponsors of German Science.] *Die Neueren Sprachen* (Frankfurt am Main), 80, 6 (1981), 541-57.

A national competition in foreign languages for pupils at school has been introduced in the Federal Republic of Germany in 1981-2 after a successful trial run in two *Länder*. In the pilot scheme, 24 out of 133 competitors reached the third round, and four girls gained first prize. Pupils must offer at least two languages and are tested in both, though at different levels. The tests are both written and oral, supervised and unsupervised; a multilingual group discussion occurs in the third round [details of tests and assessment procedures]. The aim is not, as in mathematics, to identify early brilliance, but to enhance language study (currently being criticised) at school and to stress the value in multilingual Europe of knowing more than one foreign language.

174. WILLIG, Peter

Assessment of oral performance in modern languages at Ordinary Level. *Modern Languages* (London), 63, 4 (1982), 237-46.

Complex problems are involved in the assessment of oral performance, previously the subject of vague, impressionistic marking. Although various tests have been devised to quantify specific areas of oral performance, the most effective (in the hand of a skilled examiner) is still the 'general conversation' or question/answer sequence. It is essentially flexible, probing and wide-ranging, and can develop according to an individual candidate's needs. The oral examination is unique in that it is conducted on a one-to-one basis and interactive elements play a vital role. The examiner is not only diagnostically testing the candidate but simultaneously assessing the performance. Assessment criteria must lend themselves to quick, reliable and consistent application as well as being theoretically sound.

Three broad categories of oral performance skills are: (a) the receptive, i.e. comprehension, (b) the active, subdivided into response, accuracy and range, and (c) the mechanistic (production and intonation). A cross-reference grid can be devised which reflects the performance of individual candidates and a profile can be drawn up. [Use of a picture for diagnosing oral proficiency, both with an 'open approach' and with a set of graded structure-specific questions.] Role-play is best suited to A-level or beyond. A new device is a data sheet (e.g. about activities on a particular day) which candidates prepare beforehand and then relate as a continuous narrative.

175. WOODS, Anthony, and Rosemary Baker

Item response theory. *Language Testing* (London), 2, 2 (1985), 117-40.

This is a highly algebraic introduction to the concepts underlying item response theory supported by tables of examples from actual language tests. Despite earlier criticism, the analysis of test items via the Rasch Model can complement classical methods and provide insights otherwise not easily obtained since it shows difficulty independently of testee ability, yet measures both on the same scale.

Estimates of difficulty of different items in a test are positively correlated with each other and will depend on the abilities of the subjects who happen to appear in the sample. Item difficulty is a transitive property in that, if all items are assessing the same unidimensional skill, then theoretically an individual who answers

one correctly has that skill and therefore should answer the others correctly. This never happens in practice, suggesting that the skill is contaminated or a random distraction has influenced the result.

The Rasch Model, however, states that the response of a particular person in a sample to a particular item in a test can be expressed as a function of subject ability and item difficulty when random error in both has been discounted. The Rasch Model implies that a subject whose ability is exactly equal to the difficulty of a given item will have the same probability of answering the item correctly or incorrectly. Equivalently, the difficulty of an item can be defined as the ability of a subject who will have equal probabilities of passing or failing the item. The model assumes that, however able subjects may be, they are never certain to give the correct answer even to an easy item. The stability of difference between item difficulties on the Rasch Model facilitates the banking of items whose relative difficulties are known without the need to test all the items in the bank simultaneously. Once a bank of items has been established, the apparent ability of a subject will not depend on which items are included in his test, though the standard error will depend both on the number of items and on their difficulties.

176. ZHENG CHEN, and Grant Henning
Linguistic and cultural bias in language proficiency tests. *Language Testing* (London), 2, 2 (1985), 155-63.

This study examined the English as a Second Language Placement Examination (ESLPE) employed at the University of California, to determine the nature, direction and extent of bias present for members of two linguistically and culturally diverse subgroups of the sample of examinees. By comparison of the response patterns of 34 native speakers of Spanish and 77 native speakers of Chinese from among a total sample of 312 students tested with one form of the test, it was possible to identify test items exhibiting bias in their respective skill domains. Included is a discussion of the nature, direction, extent and implications of the bias detected.

Only four vocabulary recognition items out of 150 were found to exhibit bias. These consisted of English words for which close cognate forms existed in Spanish but not in Chinese. It was concluded that as long as lexical items in the test are randomly and broadly sampled, and do not exceed the proportion existing naturally in the languages in question, there is no need for concern about fairness, even though Spanish speakers have a natural advantage.

LATE ENTRIES. Abstracts from LANGUAGE TEACHING 1987.

177. ANTONY, Kunnan John

Making classroom testing useful to teachers and learners. *Indian Journal of Applied Linguistics* (New Delhi), 12, 1/2 (1986), 101-17.

This paper maintains that classroom testing can provide vital feedback on syllabus design and teaching methodology, in such a way that educational programmes can be modified. An evaluation case study is outlined, which involved the use of a battery of formal and informal test components on 42 ESL secondary school students in the Indian state of Karnataka. Formal test items included comprehension, cloze and letter writing exercises, whereas informal assessment was undertaken by means of oral description.

The precise nature of each test element is explained, and it is noted that the guided interview was assessed with reference to a five-point oral proficiency scale, based on Foreign Service Institute specifications [tabular data]. It is concluded that school systems should recognise the value of classroom diagnostic tests, and provide flexible curricula which can allow the instructional fine-tuning which could potentially be called for by test results.

178. BACHMAN, Lyle, and Sandra J Savignon

The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview. *Modern Language Journal* (Madison, Wis), 70, 4 (1986), 380-90.

The authors believe that the Proficiency Guidelines could be elevated by enthusiasts into a classroom panacea, and call instead for the development of 'common metric tests' that could measure a wide range of linguistic abilities across different languages and in diverse contexts. Using the hybrid concept of Communicative Language Proficiency (CLP), they outline problems in such areas as the notion of direct tests and the interpretation of criterion referencing, particularly as 'zero' or 'complete' proficiency ratings elude definition and ability domains are difficult to categorise. Various drawbacks in the ACTFL oral scales are discussed, most notably the idea of the 'educated native speaker' norm and the possible interference of Test Method Factors on the spoken discourse actually elicited. Scale definitions [examples] basically seem to be context dependent and aimed *de facto* at LSP situations. The supposed empirical bases of the Guidelines are questioned as is their seeming discrete point bias. It is finally asserted that highly specific language-use CLP definitions cannot provide common metric testing

methods, and the authors propose a more generalised, global scheme. It is also felt that the Guidelines should be reviewed to take account of such factors as grammatical, discourse and sociolinguistic competence.

179. CURRALL, Steven C, and Roger E Kirk
Predicting success in intensive foreign language courses. *Modern Language Journal* (Madison, Wis), 70, 2 (1986), 107-13.

This article begins by reviewing traditional methods of placing students in appropriate foreign language courses, which mostly involve the use of standardised tests and scrutiny of predictor variables such as general intelligence, attitude and motivation. It is felt that aptitude batteries such as the MLAT and PLAB have an uneven predictive track record, and are in any case not practical evaluative tools for classroom applications.

The authors describe a project wherein 46 participants on a college intensive French course were scrutinised in terms of ten predictor variables (e.g. sex, reason for taking the course, experience of previous French courses), by means of interviews, tests and continuous assessment [tabular data]. It is inferred from the findings that overall college GPA (Grade Point Average) is the best indicator of success, though previous L2 exposure as well as overall motivation also has a discernible influence. The use of such readily defined predictors is a genuine alternative to standardised test batteries.

180. D'ANGLEJAN, Alison, and others
Beyond the language classroom: a study of communicative abilities in adult immigrants following intensive instruction. *TESOL Quarterly* (Washington DC), 20, 2 (1986), 185-205.

The study reported in this article examined the ability of adult immigrants to communicate in an interview situation as they were completing a 30-week, 900-hour second-language programme and then six months later. Foreign Service Institute (FSI)-type interviews were carried out with two cohorts of learners of varying levels of ability. The first cohort was comprised of immigrants from Southeast Asia; the second included Poles and Latin Americans. Multivariate statistical comparisons of the FSI levels at the time of the first and second interviews showed significant gains for both cohorts. However, the two cohorts differed on the subtests which contributed to those gains. Progress by Cohort 1 was mainly attributable to vocabulary gains, while gains in vocabulary, grammar, and comprehension contributed significantly to the progress made by Cohort 2. Descriptive analyses of results of subjects who progressed, regressed, or remained stable over the six-

month period showed a greater tendency to progress among lower-level subjects than among more advanced subjects.

181. DE BOT, Kees, and others

Het meten van culturele orientatie in relatie met taalvaardigheid. [The measurement of cultural orientation in relation to language proficiency.] *Toegepaste Taalwetenschap in Artikelen* (Amsterdam), 22, 2 (1985), 33-49.

Recently, numerous studies evaluating bilingual education programmes have shown that cultural orientation and language attitude play an important role in achieving sufficient proficiency in both the first and second language. Despite a long tradition of research in the field of attitude measurement, no test measuring cultural orientation in young second language learners is readily available. In this article a description is given of the development of such a test for Turkish and Moroccan children learning Dutch as a second language in primary school. The test is based on Zirkel and Jackson's (1974) Cultural Attitude Scales.

The reliability and validity of the test have been investigated in two studies. In the first (Buster *et al.*, 1985) 850 children were tested, and their scores on the cultural attitudes scales were compared to other measures that might reflect cultural orientation (questionnaires, sociograms) and to language proficiency tests. The reliability appeared to be fairly high, but there was no significant relation between cultural attitudes and language proficiency.

In the second study (Verhoeven & Extra, 1983) cultural attitude scales were used with individuals rather than with groups. Test scores were compared with data from interviews with the children tested, their parents and their teachers. On the whole this study suggests that the validity of the cultural attitude scale is sufficiently high. If certain requirements (e.g. regular revision because of the changing socio-cultural setting) are met, the cultural attitude scales appear to be a useful instrument in bilingual education research.

182. DOUGLAS, Dan

From school to university: language policy and performance at the University of Khartoum. *International Journal of the Sociology of Language* (Amsterdam), 61 (1986), 89-112.

This article focuses upon the linguistic resources of students at the University of Khartoum, considering amongst other things the relationship between Arabic/English language proficiency and academic performance. The article is based on the Study Habits Research Project completed by the author in 1977. Language policies

at the University and in the Sudan generally are outlined, and reference made to the differences between both school and university academic standards and the 'official' and 'practical' learning situations operating within the university itself. The author describes a test battery (consisting of Arabic/English reading and dictation elements) administered across the Arts, Economics, Law and Science faculties. Cloze passages represented a significant part of the battery. In general the results [tabular data] highlighted such points as the fact that most humanities students were unable to comprehend what they read in English and that women in the science faculty seemed less proficient than men in reading ability. Also, differences in Arabic test scores were felt to suggest real differences in academic performance terms between rural and urban students.

It is concluded that this linguistic assessment revealed problems in the divergent educational/language policies at the secondary and university levels and in different regions; in fact, native staff and students were becoming less proficient in English because of these inconsistencies, with a deleterious effect on the academic work of the University.

183. DUCROQUET, Lucile

Practical problems for communicative and language testing. *British Journal of Language Teaching*, 24, 3 (1986), 147-51.

In testing terms, most language teaching professionals have adapted the communicative approach without following it to its logical conclusion; the author's comments are based on administering APLL French tests to 60 students at 30 schools in Bristol, England. Public tests, even if purportedly communicative, involve such artificial methods as discussions in the L2 between people having the same L1 or who must pretend the existence of a biographical information gap (i.e. teachers might ask students questions to which they already know the answers). The asking of inordinately personal questions (which will never arise in real life) is another pitfall, as is the fact that actual students have activities and lives that do not fit the 'textbook' formulations provided in classrooms.

A wider interpretation of 'communicative competence' is called for, which would allow message transmissions by both linguistic and non-linguistic means. Communicatively appropriate (if not textbook perfect) alternatives should also be acceptable and literal translations discouraged. The use of prompt pictures which are of a

high graphical quality and unambiguous in content is also called for by the author.

184. FLYNN, Suzanne

Production vs. comprehension: differences in underlying competences. *Studies in Second Language Acquisition* (Bloomington, Ind), 8, 2 (1986), 135-64.

In this paper empirical data are presented that demonstrate significant differences in the manner in which two commonly used experimental tasks evaluate developing linguistic competence in adult second language (L2) learners. Results indicate that production tests principally evaluate a learner's developing structural competence in the L2. On the other hand, comprehension tests provide a less direct measure of structural competence and are significantly influenced by pragmatic context.

Fifty-one adult Spanish speakers at three levels of ESL ability were tested in both their elicited imitation (production) and act-out (comprehension) of complex sentences that were structurally identical. The stimuli varied in terms of the pre- and post-posing of a subordinate adverbial *when* clause and in terms of the direction of pronoun anaphora (forward and backward). Results from the production test indicate a significant effect of directionality, i.e. a preference for forward pronouns in post-posed clauses at the intermediate level. Results from the comprehension test did not show a significant directionality constraint at any level. They did, however, demonstrate a significant effect due to the use of a biasing pragmatic lead. Such findings are comparable to those found in early first-language acquisition (Lust, Loveland & Kornet, 1980). Implications of these findings for experimental methodology are discussed.

185. HANNON, Peter, and Joe McNally

Children's understanding and cultural factors in reading test performance. *Educational Review* (Birmingham), 38, 3 (1986), 237-46.

An investigation was carried out into the relationship between children's performance on a multiple-choice sentence-completion reading test and their understanding of the test items. Three groups of 24 primary school pupils (aged 7;10 to 8;11) from working-class homes, middle-class homes, and from homes where English was a second language were studied. Their understanding of test items, determined by administering the reading test orally, was found to vary considerably according to cultural background, and to be partly

responsible for social class differences in test performance. An examination of the test items suggested some reasons why this might be so: the superior performance of middle-class children may be due to a close match between their understandings and those of the test constructor(s). Implications for educational practice and research are identified. Multiple-choice sentence-completion tests often require children to contribute meanings from their own experience and cultural backgrounds. A cloze-procedure test or a reading comprehension test which presents children with more text might be less dependent on the child's contribution. If reading tests are to be used, they must be examined to see what assumptions they make about the understandings and cultural background of the children. Criterion-referenced tests tied more closely to particular curricula may serve the purpose better than norm-referenced tests based on a 'national' culture.

186. JOYCEY, E

Teaching towards the text and the multiple choice question type test. *System* (Oxford), 15, 1 (1987), 89-95.

In basic terms, the author presents ten classroom techniques which could help students to reconcile the communicative teaching of reading with the demands imposed by the traditional multiple-choice exercises still commonly included in public examinations. Meaningful reading-as-process strategies can be developed to cope with multiple-choice questioning, so that students can infer the relationship between the alternatives in each question and the text itself (thus harnessing the potentially negative backwash of multiple-choice tasks).

The techniques presented include: allowing the learners to answer questions without seeing the text (to exploit their knowledge of the world), to reduce each multiple-choice question to a single *wh*-form and to try writing their own multiple-choice questions.

187. LITTLEWOOD, William

GCSE: some problems and opportunities. *British Journal of Language Teaching*, 24, 2 (1986), 63-9.

The GCSE offers an opportunity to integrate some important strands of development within foreign language teaching: new definitions of language teaching aims, moving towards what people can do with the language; new orientations in teaching methods - communicative activities; new expectations from assessment - evaluating ability in global terms, and assessing in positive terms;

new views about how people learn - the importance of unconscious 'acquisition' which takes place when language is used for communication, and a gradually evolving mastery of the system as learners form their own rules (in this view, errors are inevitable). In a broader educational context, the GCSE embodies modern trends such as developing learning programmes based on learners' needs and wants, assigning active learning roles, involving the whole person, and viewing knowledge as something which learners construct. The more positive approach to evaluating performance and the use of common scales for all learners, offer greater prospects of devising broadly comparable approaches for all learners. The GCSE is not a new theory, but a new framework.

The work which remains to be done involves sorting out problems in three main areas: (1) the assessment system itself, (2) the kind of teaching needed, and (3) the introduction of the new assessment and teaching into the whole of the language teaching population. Under (1), the notion of 'effective communication' or 'appropriacy' needs to be defined in operational terms which mean the same to all assessors. Norm referencing will give way over the next few years to criterion referencing. Under (2), although there is a common goal for all learners, there can be many routes to that goal. Innovations do not simply involve adjustments in techniques, but changes in aims and attitudes. Only a meagre in-service training programme is envisaged so teachers will have to try to extend the spirit of dialogue and co-operation which already exists.

188. MUNDZECK, Fritz

Normierungstendenzen im Fremdsprachenunterricht - Beispiel Textaufgabe. [Text analysis as an example of standardising tendencies in modern language teaching.] *Die Neueren Sprachen* (Frankfurt am Main, FRG), 85, 4 (1986), 405-20.

The author examines the degree of standardisation that has taken place in the teaching and testing of written language skills following the uniform examination requirements (EPA) recommended by the German *Länder*. The *Textaufgabe*, or text analysis, is one example of an exercise which has become highly standardised; it is examined (in terms of its origins, structure, didactic purpose, and prevalence both in the *Länder* guidelines and in the classroom) in order to demonstrate how standardised exercises of this type become established in all language courses and play an ever more important role in classroom work with texts. The article thus provides a critical analysis of the standardisation process and

seeks to stimulate discussion of possible alternatives to this approach.

189. PAGE, Brian

GCSE: motorway or cul-de-sac? *ATI Journal* (Beccles, Suffolk), 48 (1986), 12-19.

The National Criteria now being implemented in syllabuses for the new GCSE examinations represent a giant step forward which may come as a shock to many teachers. For the first time, public examinations have clearly stated aims (which see the language as an instrument for practical communication), an outline is given of what sort of test types are appropriate, and the candidate's final performance is to be described in terms of completing specific tasks or demonstrating particular skills. The increasing degree of central control, though essential, is worrying. The National Criteria need to be revised and kept up to date, but this is not being done. The draft Grade Criteria look like being far too rigidly constraining, which will stifle any future growth of syllabuses. Another worry is that writing is still essential for the most able pupils. The aims of the National Criteria are beginning to look limited; they say too little about the development of the individual as a person or as a social individual. The division into the four skills was accepted for lack of anything better, but it is being hardened in an unfortunate way by the Grade Criteria. To offset this, there will need to be definite encouragement of mixed-skill teaching and testing to break down the artificial divisions between skills. Another weakness in the National Criteria is their failure to do much about course assessment in foreign languages, though it figures largely in GCSE as a whole.

On the positive side, GCSE in modern languages represents a fundamental shift of emphasis in two ways: (1) it is concerned with language use and meaning, and (2) it is concerned with assessing and describing what individual candidates can do, not what they can not do, and not comparing their performance to other people's (i.e. a move from norm referencing to criterion referencing). It is essential to see the lower grades as positive attainments. Although the Grade Criteria seem inflexible, there are possibilities, such as in the school-based assessment mode, for extending it in the right direction, so long as teachers retain the initiative.

190. PERKINS, Kyle, and others

Derivational complexity and item difficulty in a sentence repetition task. *Language Learning* (Ann Arbor, Mich), 36, 2 (1986), 125-41.

This paper reports the results of an analysis to determine whether derivational complexity is a determinant of item difficulty in a sentence repetition task. Fifty ESL subjects took a 26-item sentence repetition test. It was possible to estimate Rasch item difficulty indices for 18 items. The high-difficulty items were derivationally more complex than the low-difficulty items; however, predictions made in the early derivational complexity research about passive, negative, and interrogative sentences were not confirmed in this research. A linguistic analysis of the most difficult items suggests that this sample of ESL subjects experienced difficulties in processing adverbials, compounding and reduction of clauses, and the non-finite verbals.

191. POWERS, Donald E

Academic demands related to listening skills. *Language Testing* (London), 3, 1 (1986), 1-38.

A literature review was conducted in order to identify various parameters underlying listening comprehension. The results of this review were used as a basis for a survey of faculty in six graduate fields as well as undergraduate English faculty. The purpose of the survey was to (a) obtain faculty perceptions of the importance to academic success of various listening skills and activities, (b) assess the degree to which both native and non-native speakers experience difficulties with these skills or activities, and (c) determine faculty views of alternative means of evaluating these skills.

Faculty perceived some listening skills as more important than others for academic success. These included nine skills in particular that were related primarily to various aspects of lecture content (e.g. identifying major ideas and relationships among them). As might be expected, faculty perceived that non-native students experience more difficulty than native students with all listening activities, and that non-native students have disproportionately greater difficulty with some activities, such as following lectures given at different speeds and comprehending or deducing the meaning of important vocabulary. With respect to measuring listening comprehension, some general approaches and specific item types were judged to be more appropriate than others. These included tasks that entail answering questions involving the recall of details as well as those involving inference or deductions.

The results of the survey are used to suggest further research on the construct validity of the Listening Comprehension section of the TOEFL.

192. SANG, F, and others

Models of second-language competence: a structural equation approach. *Language Testing* (London), 3, 1 (1986), 54-79.

Starting with a critical discussion of the consensus reached concerning the structure of second-language competence, the unitary competence hypothesis as advocated especially by Oller (1976) is confronted with new evidence supporting a multidimensional model of foreign-language ability. In order to question the plausibility of the general (unitary) factor hypothesis, its central claim, that of universal validity, is put to the test. This claim implies at least two assumptions: (a) the invariance of the general factor, regardless of differences in the populations studied, and (b) the invariance of this factor irrespective of the conditions under which language acquisition takes place. By showing divergent structures in subgroups of high and low first-language ability, the first assumption would be falsified. By demonstrating specific influences of two teaching styles on different language components, the second assumption could be disproved.

The analyses are based on data from a sample of 14,000 seventh-grade students and their teachers from 427 secondary schools (*Gymnasien*) being highly representative for West Germany and West Berlin. The hypotheses were tested using confirmatory factor analysis (LISREL). The results obtained show the appropriateness of the three-factor model proposed (ELEMENTARY, COMPLEX, COMMUNICATIVE), which is theoretically meaningful and best fits the data. Moreover, the expected effects of first-language ability grouping and of teaching strategies on the structure of second-language competence clearly emerge. In spite of the limitations of the data set analysed, these results can all be interpreted as strong and further evidence against a single-factor model (and its educational implications) and in favour of a multiple-factor model as outlined above. Didactic consequences of these findings are discussed.

193. SCHULZ, Renate A

From achievement through proficiency through classroom instruction: some caveats. *Modern Language Journal* (Madison, Wis), 70, 4 (1986), 373-9.

The author warns against what she sees as an extremist trend which would elevate the ACTFL Provisional Proficiency Guidelines to the status of a 'blueprint' for curriculum design, with the resulting (and hasty) imposition of oral proficiency as the central skill in

language programmes. She questions the wisdom of giving oral proficiency such a role, particularly since there is no consensus on the exact nature of language proficiency, and because oral abilities are the most time-consuming to develop and also the easiest to 'lose'. Moreover, it is felt that L2 oral fluency may not be in any case a sound pedagogic goal for US university or high school students, who might never need to interact in the foreign culture or to perform any linguistic functions apart from an occasional use of comprehension skills. Teaching situations and learner constraints also complicate the issue.

The establishment of an Acquisition methodology in formal instruction would be beneficial: the position of the author is contrasted point by point with commonly held beliefs on language learning. She concludes by calling for a broader-based FL classroom which teaches for and measures multi-skill development, and cultural/linguistic awareness.

194. SCOTT, Mary Lee

Student affective reactions to oral language tests. *Language Testing* (London), 3, 1 (1986), 99-118.

Many factors influence student reactions to tests, including format, length time constraints, testing environment, familiarity with test format, perceptions of test validity and student anxiety. The present study assessed the affective reactions of native Brazilian students to different oral EFL test formats (group and pair) in an achievement-testing situation. Each student participated in both the group and the pair test formats and completed an affect questionnaire after each test. A principal-factor analysis of student responses to the questionnaire revealed two major factors. The first appeared to involve cognitive judgements about the validity of the test and how well students liked the test. The other factor seemed to be emotive, and involved student reports of the anxiety they felt both before and after the test. Similarly, a qualitative analysis of student responses to open-ended items on the questionnaire found that students were mostly concerned about their emotional state, test content and time constraints in the testing situation, while comments about test administrators accounted for only a small percentage of the total responses. A multivariate analysis of variance based on the results of the factor analysis showed no significant difference among student reactions to the different test formats.

195. SHOHAM, Miriam, and others Reading comprehension tests: general or subject-specific? *System* (Oxford), 15, 1 (1987), 81-8.

Tests of reading comprehension have traditionally been based on tests which are considered to be comprehensible to the educated layman, thus not requiring specialist knowledge. However, recent developments in language testing are moving towards subject-specific tests. Such ESP (English for Specific Purposes) tests are based on the belief that it is more valid to test the reading comprehension of an engineering student on an engineering topic rather than on a social science topic and vice versa. Empirical evidence in favour of either the general or the specific approach is lacking. This paper is a report of a study designed to investigate the relevance of student background discipline on tests of reading comprehension in EFL (English as a Foreign Language). One hundred and eighty-five students from three faculties: Science and Technology, Biology, and Humanities and Social Science, were tested on three texts related to their respective fields of study. It was found that content area passages do affect student performance on reading comprehension tests, but not as greatly as had been expected. It was also found that the order of presentation of a reading passage in a multiple-text test does not affect student performance.

196. SHOHAMY, Elana, and others
Introducing a new comprehensive test of oral proficiency. *ELT Journal* (London), 40, 3 (1986), 212-20.

The study reported in this article attempted to experiment with and develop a new oral proficiency test which could replace the existing EFL Oral Matriculation test administered by the Ministry of Education in Israel. The deficiencies of the existing Oral Matriculation test are specified, and the components of the experimental test are described.

The analysis of the data showed that the experimental test had better linguistic, educational and testing qualities than the existing Oral Matriculation test; namely, it produced a better distribution of scores, showed reasonable rater-reliability, tested a broader range of speech styles, and produced favourable attitudes on the part of the test-takers.

The recommendation to replace the existing test by the new experimental test has since been accepted by the Ministry of Education.

197. TAKASHIMA, Hideyuki
Acculturation and second language learning: use of comics to measure the degree of acculturation. *IRAL* (Heidelberg, FRG), 25, 1 (1987), 25-40.

A study investigated whether tests based on comics or cartoons can be used as effectively and reliably as other tests such as the cloze to measure the degree of acculturation or relative proximity of a student to a target language. It was hypothesised that scores on comics would not correlate well with cloze test scores if the comics were extreme in either linguistic or cultural directions. The test format consisted of a cloze test and interpretations of four out of six cartoons.

Results showed that there was a good correlation between the cloze test and the cartoons. Students who scored higher on the cartoons could be considered more acculturated; though they may not be fluent, they are approaching the target in one way.

198. WHELPTON, Tony

The assessment of oral performance at O- and A- level. *NIMLA* (Coleraine, N. Ireland), 14/15 (1985), 29-37.

The aim of an oral examination is to assess the extent of the candidate's ability to communicate in a foreign language. Grammatical accuracy is only one of the aspects which are taken into consideration. The AEB examination is criterion-referenced, not norm-referenced. The format of the O- level oral examination was re-designed in 1975 to include role-playing, standardisation being provided by a series of pictures about which the candidate had to relate the story. The response of the examiner in the role-play is not known beforehand. The longest section of the exam consists of general questions, testing ability to sustain a conversation. The examiner is interested in what the candidate produces, not what he/she has understood. The examiner tries to find out what each candidate can do - i.e. his/her linguistic ceiling. The marking is positive - a flawless performance is not expected at this level.

199. YOUNG, Dolly J

The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals* (New York), 19, 5 (1986), 439-45.

The American Council on the teaching of Foreign Languages (ACTFL) has developed an oral proficiency test, the Oral Proficiency Interview (OPI), which may become the official test of oral performance for prospective foreign language teachers in Texas. Professionals in foreign language education are raising concerns

about the effect of anxiety on Oral Proficiency Interview ratings. The purpose of this study was to provide an assessment of how anxiety may influence scores on the OPI.

An examination of the relationship between anxiety and oral performance must take foreign language ability into account. In this study, subjects' foreign language proficiency was assessed through a Self-Appraisal of Language Proficiency questionnaire and a dictation test. Subjects' anxiety was assessed on the basis of four independent anxiety measures: the State Anxiety Inventory, the Cognitive Interference Questionnaire, A Self-Report of Anxiety, and a Foreign Language Anxiety Scale of Reactions.

A total of 60 subjects took the OPI at various American universities. These subjects were either majors in French, German, or Spanish or prospective French, German, or Spanish teachers at one of these three institutions. The Self-Appraisal of Speaking Proficiency and the State Anxiety Inventory were administered before the OPI, and the other anxiety instruments were given immediately after it.

Results showed significant negative correlations between anxiety and the OPI, but once the effects of ability were accounted for, the correlations were no longer significant. In other words, once the effect of an individual's language proficiency was accounted for, oral performance no longer decreased as anxiety increased.

An important qualification that must be noted is that subjects were aware that the OPI in this study represented an unofficial administration of the test. Therefore, this study does not provide data on anxiety levels for official test situations.

200. ZEIDNER, Moshe

Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing* (London), 3, 1 (1986), 80-98.

The major aim of this study was to examine the validity and crosscultural generalisability of the test bias contention as applied to English Language aptitude test, routinely employed for students election and placement functions within the Israeli academic scene. The bias analysis was based on the English language aptitude test scores of $n = 1778$ Jewish and $n = 1017$ Arab student candidates applying for admission to a major campus in northern Israel, who were administered the test as part of the scholastic aptitude test battery used for college admissions purposes. The psychometric properties of the English language aptitude test were compared for

ABSTRACTS

Jewish and Arab subgroups, respectively, via a variety of internal (e.g. factor loading, reliability, standard error of measurement, etc.) and external (e.g. predictive validity, homogeneity of regression, standard error of estimate, etc.) criteria. The test scores show significant but slight intercept bias, tending to overpredict the overall first-year GPA of Arab student candidates. Also, test scores appear to be somewhat less reliable measures for Arab compared to Jewish student candidates. On the whole, however, our data provide only marginal evidence for differential construct or predictive validity of English language aptitude test scores as a function of cultural group membership, thus lending a greater amount of generality and crosscultural validity to the findings of much previous research negating the cultural bias hypotheses.

3. Survey articles

S1. DAVIES, Alan

Language testing - Part I. Survey article in *Language Teaching & Linguistics: Abstracts*, Vol 11, No 3, July 1978. pp 145-59.

Language testing - Part II.

Survey article in *Language Teaching & Linguistics: Abstracts*, Vol 11, No 4, October 1978. pp 215-31.

In Part I the author discusses discrete-point tests with particular reference to the major textbooks on language testing. Then he goes on to a brief discussion of tests of reading comprehension and ends with a note on criterion reference testing. In Part II he discusses integrative tests, giving special consideration to cloze tests and dictation tests, followed by a discussion of tests of spoken and written production. Part 2 ends with a consideration of the testing of communicative competence. A full bibliography is given for both parts at the end of Part 2.

S2. DIRVEN, R, and J Oakeshott-Taylor

Listening comprehension. State of the art survey article. Part 2 in *Language Teaching*, Vol 18, No 1, 1985. Section: The testing of listening comprehension. pp 15-16.

S3. PAGE, B

Graded objectives in modern-language learning. Survey article in *Language Teaching*, Vol 16, No 4, October 1983. pp 292-308.

Main sections headed: The beginnings; The growth of the movement; The syllabuses; Tests; Materials and methodology; National co-operation; International interest; Misgivings; Pupil and teacher motivation; Conclusion. Extensive bibliography.

S4. SKEHAN, P

Testing - Part 1. Survey article in *Language Teaching*, Vol 21, No 4, October 1988 (Forthcoming).

Testing - Part II. Survey article in *Language Teaching*, Vol 22, No 1, January 1989 (Forthcoming).

4. Bibliography

4.1. SOURCES OF INFORMATION ON TESTING AND EXAMINATIONS

B1. ALDERSON, J C, and others, editors

Reviews of English language proficiency tests. Washington DC: TESOL, 1987.

Provides descriptive and evaluative information on the major ESL/EFL tests being used in the world today. Reviews contain a description of the test, a discussion of its reliability, validity, and related issues. Also included are bibliographical references which provide further information on each test.

B2. BUROS, Oscar K, editor

Tests in print 2. Highland Park, N.J.: The Gryphon Press, 1974.

An index to tests, test reviews, and the literature on specific tests.

B3. BUROS, Oscar K, editor

Foreign language tests and reviews. Highland Park, N.J.: The Gryphon Press, 1975.

Consists of the Foreign Language Sections of the *Mental measurements yearbooks*, 1-7, and *Tests in print*, 2.

B4. BUROS, Oscar K, editor

Mental measurements yearbooks. Vols 1-8. Highland Park, N.J.: The Gryphon Press, 1938-1978.

B5. *CILT Information* (Available free from CILT. Please send A4 SAE).

A series of short information sheets. Revised, and new titles prepared from time to time. 10: Patterns of change in GCE 'A' Level examinations. 40: Guide to GCSE: French. 51: Guide to GCE 'AS' level examinations. 57: Profiling, coursework and continuous assessment: a list of references.

B6. *EDUCATION Year Book.* Harlow: Longman. Annual.

The names and addresses of all the British examining boards are listed. Enquiries about syllabuses and past papers should be directed to the Examination Board concerned.

BIBLIOGRAPHY

- B7. LONGMAN Exam Guides Series, 1986-88.

English as a foreign language: beginner examinations. Published 1987.

English as a foreign language: intermediate examinations, by F. O'Dell. Published 1986.

English as a foreign language: advanced examinations. Published 1988.

French examinations, by T Carter. Published 1987.

These guides describe all the examinations available, thus helping the student to assess which one is suitable. Each guide contains information on major course topics and gives advice on studying, revision and examination technique.

- B8. *The PITMAN guide to English language examinations.* 2nd ed. London: Pitman, 1984.

A comprehensive guide to UK-based English language examinations for non-native speakers. Provides full descriptions and specimen question papers for the examinations of 18 boards.

- B9. SAVARD, J G

Analytical bibliography of language tests. Québec: Presses de l'Université Laval (for) International Center for Research on Bilingualism, 1969.

Brief description of a variety of second-language, mother-tongue and bilingual tests in English, French, German, Italian, Spanish. (Some unpublished and many now out of date).

- B10. SAVARD, J G

Analytical bibliography of language tests. 2nd ed. rev. and enlarged, 1977.

Continues the first edition. 156 second-language, and 244 mother-tongue tests, covering Afrikaans, Chinese, Dutch, English, French, German, Hebrew, Italian, Latin, Russian, Spanish.

- B11. SIMMONDS, Paul

A survey of English language examinations. *ELT Journal*, 39, 1, 1985. pp 33-42. See Abstract 155.

4.2 BOOKS (MONOGRAPHS, CONFERENCES, SURVEYS, RESEARCH REPORTS)

- B12. ALDERSON, J C

Innovation in testing: can the micro-computer help? Special edition

of *Language Testing Update*. Lancaster: University of Lancaster, Institute of English Language Education, 1987.

Report of research into computer-based English language testing, commissioned 1985 by the British Council and carried out at Lancaster.

B13. AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGES

The ACTFL provisional proficiency guidelines. These were first published by ACTFL in 1982. Further developments are reported in the following sources (among others not cited)

BYRNES, H, and M Canale, editors

Defining and developing proficiency: guidelines, implementations, and concepts. Lincolnwood, Ill.: National Textbook Co., 1987.

HIPLE, D V

A progressive report on the ACTFL proficiency guidelines, 1982-1986. Lincolnwood, IL.: National Textbook Co., 1986.

The ACTFL journal *Foreign Language Annals*.

Language-specific sets of proficiency statements currently exist for French, German, Spanish, Chinese, Japanese and Russian. Further information from ACTFL, 579 Broadway, Hastings-on-Hudson, New York, USA 10706.

B14. ASHWORTH, A E

Testing for continuous assessment. London: Evans, 1982.

For practising teachers in schools and colleges. Explains the basic ideas in testing, measurement and evaluation in the classroom.

B15. BACHMAN, L F

Basic concerns in language testing validation. Reading, Mass: Addison-Wesley, 1989 (forthcoming).

B16. BIRD, E, and M Dennison

Teaching GCSE modern languages. London: Hodder & Stoughton, 1987.

Describes the national criteria for modern languages, the details of the GCSE examination structure, the types of assessment and

differences among the various syllabuses on offer, and the implications for classroom practice.

BLACK, P.J, Chairman. *See* B57.

B17. BOLTON, S

Die Gütebestimmung kommunikativer Tests. Tübingen: Gunter Narr Verlag, 1985.

L'évaluation de la compétence communicative en langue étrangère. Paris: Hatier-Credif, 1987.

Describes the evolution of tests of communicative competence and the influence on them of current linguistic theory from structuralism to pragmatics.

B18. BROADBENT, J, and others

Assessment in a multicultural society: community languages at 16+. York: Longman Resources Unit, for Schools Council, 1983.

Examination provision in Urdu, Hindi, and Modern Greek is explored, and proposals for a new framework for assessment in each language are made.

B19. BROADFOOT, Patricia, editor

Profiles and records of achievement. London; New York: Holt, Rinehart, 1986.

Reviews initiatives in the field of profiles and records of achievement and progress made by various bodies, including LEA's, examination boards, Graded Objectives Movement, TVEI scheme, etc.

B20. BROADFOOT, Patricia

Introducing profiling. Basingstoke: Macmillan Education, 1987.

Analyses schemes already developed, and considers practical issues involved in their implementation. Short, but useful bibliography. Dr Broadfoot is Co-Director of the National Evaluation of the DES-funded Pilot Projects in Records of Achievement.

B21. BUCKBY, M, and others

Graded objectives and tests for modern languages: an evaluation. London: The Schools Council; distributed by CILT, 1980.

Useful, achievable short-term objectives can transform the attitudes of parents and children towards the learning of French, according to this report of the Schools Council Project on Graded Tests in Modern Languages.

BYRNES, H, editor. See B13.

B22. CARROLL, B J

Testing communicative performance: an interim study. Oxford: Pergamon, 1980.

A study of the influence of communicative syllabus design on language testing. The author analyses the British Council's work on performance testing based on experience with foreign students in the UK, East Africa, South America and the Middle East.

B23. CARROLL, B J, and P Hall

Make your own language tests: a practical guide to writing language performance tests. Oxford: Pergamon, 1985.

'This handbook provides practical guidance to teachers on the design, development and use of communicative language tests to assess not just knowledge of a language system but also the ability to use language in real situations'. - Publisher's catalogue. Appendix 3: Annotated bibliography. Appendix 4: A guide to selected communicative language tests.

B24. COHEN, Andrew D

Testing language ability in the classroom. Rowley, Mass: Newbury House Publishers, 1980.

Emphasis on testing while teaching, on day-to-day evaluation, and on assessing communicative skills. The suggested items and tests presented are taken from the English language.

B25. COUNCIL OF EUROPE. Modern Languages Project.

Approaches to self-assessment in foreign language learning, by M Oskarsson. New ed. Oxford: Pergamon Press for and on behalf of the Council of Europe Modern Languages Project, 1980.

An outline of possible forms of guided self-assessment in adult language learning, review of research and practice in a number of countries and suggestions for materials and their use.

B26. COUNCIL OF EUROPE. Modern Languages Project.

Self-assessment of foreign language skills: a survey of research and development work, by Mats Oskarsson. Strasbourg, 1984.

Thirteen pieces of research are reviewed, four of which focus on work on continuous assessment. Attention is on adult learning and on procedures and results that are judged to be of general interest. Various types of materials that have been used for self-assessment purposes are illustrated.

B27. COUNCIL OF EUROPE. Modern Languages Project.

International workshops for trainers of modern language teachers. 1984, 7b: *Task-based assessment of communicative competence.* Workshop held Bournemouth, UK, 1984. Report by F. Heyworth. Strasbourg, 1986. (CC-GP12(84)W.7b, English). 1985, 4: *The assessment and testing of communicative performance in foreign languages in secondary schools (with special emphasis on teacher training).* Workshop held York, UK, 1985. Report by P. Green. Strasbourg, 1987. (CC-GP12(85)W.4, English).

B28. COUNCIL OF EUROPE. Modern Languages Project.

Communicative language testing: a resource handbook for teacher trainers; edited by Peter Green. Strasbourg, 1987.

An edited and revised version of contributions, with supporting materials, to Workshop 1985, 4 (above). It is offered as a resource for teacher trainers, as a basis for in-service courses on the testing of the communicative proficiency of pupils in secondary schools.

B29. COUNCIL OF EUROPE. Modern Languages Project.

Evaluation and testing in the learning and teaching of languages for communication. Strasbourg, 1988.

Report consisting of papers presented by experts at two meetings. Includes papers by Samuel, Buckby, van Weeren, Kohonen, Schratz, Oskarsson, Schärer, and Mairesse.

B30. DEPARTMENT OF EDUCATION AND SCIENCE.

Assessment of Performance Unit.

The publications listed below relate to a research project commissioned by the DES and conducted by the NFER. Project Leader: Peter Dickson. Begun in 1981 the aim of the project was to conduct on a national scale a series of three annual surveys (1983, 1984, 1985) of the performance of 13-year-old pupils in the first foreign language in schools in England, Wales, and Northern Ireland. The survey reports provide information on pupils' performance in listening, reading, speaking, and writing as well as the description of the assessment materials. They give information on the schools surveyed, the pupils' background and their attitudes to foreign language learning, the foreign language department and the organization of foreign language teaching in the schools concerned. Further publications associated with APU are anticipated from NFER/Nelson - these include assessment materials, training materials, and a report, work on which began January 1987 - See also DICKSON, B34 and Abstracts 68 and 129.

BIBLIOGRAPHY

Survey reports

Foreign language. DES, 1980. Rationale of the project.

Foreign language provision. DES, 1983. (APU Occasional Paper, 2).

Foreign language performance in schools: Report on the 1983 survey of French, German, and Spanish. DES, Welsh Office, and DENI, 1985.

Foreign language performance in schools: Report on the 1984 survey of French. DES, 1986.

Foreign language performance in schools: Report on the 1985 survey of French. HMSO, 1987.

APU Newsletter. No. 7, Spring 1985.

Reports for teachers

Assessing foreign languages: the French, German and Spanish tests. Windsor: NFER/Nelson, 1986.

Foreign languages: listening and reading. NFER/Nelson, 1987. (Book and cassette).

Foreign languages: speaking. NFER/Nelson, 1987. (Book and cassette).

Foreign languages: writing. NFER/Nelson, 1987.

B31. DEPARTMENT OF EDUCATION AND SCIENCE. Her Majesty's Inspectorate of Schools.

A survey of the use of graded tests of defined objectives and their effect on the teaching and learning of modern languages in the County of Oxfordshire, carried out 17-21 May and 21-25 June 1982. Report by HM Inspectors. Published July 1983 by the DES and available from the DES Publications Despatch Centre, Honeypot Lane, Stanmore, Middx. Other similar documents since produced are listed in PAGE and HEWETT (B66) and available from address above.

B32. DEPARTMENT OF EDUCATION AND SCIENCE, and Welsh Office *Records of achievement*. Published by HMSO, July 1984.

B33. DICKINSON, Leslie
Self-instruction in language learning. Cambridge University Press, 1987. Includes bibliography. Ch. 8: Self-assessment.

B34. DICKSON, Peter (Project Leader)
Modern languages in schools (MODLS): the development of APU assessment materials for teachers.

A development research project sponsored by the DES and based at the NFER. In the course of the APU monitoring programme a wide range of assessment materials and procedures was produced. The MODLS project was established to exploit these materials. The project team is working in five LEAs - two groups are concerned with the assessment of French, two of German, and one of Spanish. Publications will include assessment materials and in-service training materials. Duration of project: January 1987 - December 1988. (Further details from NFER, The Mere, Upton Park, Slough, Berks). See also B30 above.

B35. EAST MIDLANDS GRADED ASSESSMENT
 FEASIBILITY STUDY

French for communication: the East Midlands Graded Assessment Feasibility Study, edited by R Dunning. Leicester: University of Leicester School of Education, 1983.

An account of the co-operation between the East Midland Regional Examinations (EMREB), the University of Leicester School of Education, the East Midlands LEAs and East Midlands teachers of modern languages. The book deals with Level One. Various reports have emanated from the EMREB Research Project.

B36. ELT Documents, III
Issues in language testing. London: The British Council, 1981.

Papers from a Symposium held at the University of Lancaster, October 1980. Section 1: Communicative language testing. Section 2: Testing of English for specific purposes. Section 3: General language proficiency.

B37. FALLOWS, I.
Assessment in a multicultural society: English at 16+. York: Longman Resources Unit, for Schools Council, 1983.

A discussion document reviewing GCE 'O' level and CSE syllabuses and examinations in English, including ESL, in the context of needs within multicultural education.

BIBLIOGRAPHY

B38. FRITH, DS, and H G Macintosh

A teacher's guide to assessment. Cheltenham: Stanley Thornes, 1984.

Rather technical manual of interest to teachers, teacher trainers, advisers, etc. Can be used for self-teaching. Index and glossary of terms.

B39. GARFORTH, D, and H Mackintosh

Profiling: a user's manual. Cheltenham: Thornes, 1985.

A self-help introduction to profiles and achievement for those without any formal specialist training in assessment.

GREEN, P. See B28.

B40. HARRISON, A

Review of graded tests. London: Methuen Education, 1982.

This review describes graded test systems in sport, music, mathematics, science and other areas as well as in languages.

B41. HARRISON, A

A language testing handbook. Basingstoke: Macmillan Education, 1983.

Discusses simply and clearly the different types of test: placement, diagnostic, achievement, and proficiency. Provides guidelines on adapting the most suitable kind of test to fit a particular situation, with two final chapters on marking and interpreting students' scores. Examples from English.

B42. HARRISON, A

Profile reporting of examination results. London: Methuen Education, 1983.

A report, commissioned by the Schools Council, which reviews the issues and evaluates the feasibility of profiling.

B43. HAUPTMAN, P C, R LeBlanc, and M B Wesche, editors

Second language performance testing: papers from a conference held at the University of Ottawa, March 1983. University of Ottawa Press, 1986.

Developments in the field of performance tests in Canada and elsewhere which are used for testing older learners in language training programmes and in certification testing for specific occupational situations.

BIBLIOGRAPHY

B44. HEATON, J B, editor

Language testing. Oxford: Modern English Publications, 1982.

A collection of articles written by testing specialists, teachers and teacher trainers involved in the testing of English in a number of different countries. Testing is re-examined in the light of developments in communicative language teaching.

B45. HEATON, J B

Writing English language tests. New ed. Harlow: Longman, 1988. First pub'd 1975.

Examines different types of language tests, and shows how they can be applied to different language areas. A manual for teachers who construct tests.

B46. HENNING, G

A guide to language testing: development, evaluation, research. New York: Newbury House (distributed by Harper & Row, UK), 1987.

A technical and comprehensive guide to the principles of test and questionnaire development for use in placement, achievement, evaluation and research. Includes a thorough treatment of the estimation of reliability and validity, offering a variety of techniques to fit many measurement contests. Includes computational material and formulae.

B47. HIGGINSON, G R, Chairman

Advancing 'A' levels. Report of a Committee appointed by the Secretary of State for Education and Science and the Secretary of State for Wales. London: HMSO, 1988.

The task of the Committee, set up in March 1987, was to recommend the principles that should govern 'A' level syllabuses and their assessment.

HIPLE, D V. See B13.

B48. HUGHES, A, and D Porter, editors

Current developments in language testing. London: Academic Press, 1983.

Selected papers from a seminar held under the auspices of the British Association for Applied Linguistics, at Reading University in December 1981. The three themes of the seminar were: 'The structure of language proficiency'; 'Communicative language testing'; and 'Testing language for specific purposes'.

BIBLIOGRAPHY

B49. INTERNATIONAL SYMPOSIUM ON LANGUAGE TESTING.

Hong Kong, 1982. New directions in language testing: papers; edited by Y P Lee (and others). Oxford: Pergamon, 1982.

A selection of papers presented under the headings: 'Large-scale testing'; 'Testing oral proficiency'; 'Validating language tests'. Bibliography.

B50. INTERUNIVERSITÄRE SPRACHTESTGRUPPE.

International Language Testing Symposia.

Practice and problems in language testing.

1st, Hurth, 1979. Proceedings held at the Bundessprachenamt, Hurth, 29-31 July 1979. Frankfurt/Main; Berne: Verlag Peter Lang, 1981.

3rd, Brussels, 1980. Proceedings held at the Instituut voor Talonderwijs, Vrije Universiteit Brussels, 5-7 November 1980. Brussels: Vrije Universiteit, 1982.

4th, Colchester, 1981. Proceedings held at the University of Essex, 14-17 September 1981. Colchester: University of Essex, Department of Language and Linguistics, 1982 (Occasional Papers, 26).

5th, Arnhem, 1982. Proceedings held at CITO. Arnhem, 1982. Arnhem: CITO (Dutch National Institute for Educational Measurement), 1982. (Editor J van Weeren). Out of print. Available as an ERIC Document - Accession No. ED 282 415 (108 pages).

7th, Colchester, 1983. Proceedings held at the University of Essex, 14-17 September 1981. Colchester: University of Essex, Department of Language and Linguistics, 1985. (Occasional papers, 29).

8th, Tampere, 1984. Proceedings held at the University of Tampere, November 17-18, 1984. Distributed by Akateeminen Kirjakauppa, Helsinki. Pub'd 1985.

9th, Windsor, 1985. *Innovations in language testing*. Proceedings held at the NFER, Windsor, April 1985. Windsor: NFER/Nelson, 1986.

(Publication details of the 2nd and 6th symposia not known).

B51. McEWEN, A, and others

Graded objectives in French: attitudes in Northern Ireland schools. Belfast: School of Education, Queen's University of Belfast, 1988.

BIBLIOGRAPHY

A research report funded by the Leverhulme Trust. This large-scale project took place over two years and the report covers the design of the research, comparison of experimental and control groups, teacher and pupil interviews. The interviews and the responses to the three languages attitude scales provide a considerable body of evidence that experience of graded objectives syllabuses and tests engenders more positive reactions to learning French than does the more traditional approach.

B52. MADSEN, H S

Techniques in testing. Oxford University Press, 1983.

A practical book for ESL/EFL teachers, introducing within each chapter a variety of techniques and a number of activities for the classroom. Part 1 deals with the subskills: vocabulary, grammar and pronunciation; Part 2 with the communication skills: reading, writing, listening, and speaking.

B53. MARSHALL, G

An investigation into the present state of language proficiency tests and the possible use and development of communicative tests for English as a second language. Coventry: Coventry Education Authority, Minority Group Support Service, 1981.

Report of research conducted in Coventry. Includes guidelines for drawing up assessment procedures for communicative proficiency tests. Refers to published ESL tests - Pinpoint (Leicester Language Centre); Tests of Proficiency in English (NFER); and CSE Modes I, II, III, and University Boards.

B54. MORROW, K

Techniques of evaluation for a notional syllabus. Orpington: Royal Society of Arts, 1977.

Study commissioned by the RSA.

B55. NATIONAL CONGRESS ON LANGUAGES IN

EDUCATION, 3rd Assembly, Nottingham, 1982.

Foreign language examinations: the 16+ debate 1981-83; edited by Alan Moys. London: CILT, 1984. (NCLE papers and reports, 5).

Edited papers from a working party for the third NCLE Assembly. A survey of the period of debate and reconsideration of criteria for 16+ examinations. Fully documented by the inclusion of the major statements from public bodies and associations, e.g. the GCE and CSE Boards' Joint Council, the Schools Council, and the NCLE.

BIBLIOGRAPHY

B56. NATIONAL CONGRESS ON LANGUAGES IN EDUCATION. 4th Assembly, York, 1984.

Language and languages 16-19; edited by A G Limb. London: CILT, 1986. (NCLE papers and reports, 7).

Edited papers from a working party for the fourth Assembly. They provide a survey of language education within the 16-19 age group in the light of the changes in provision (YTS, TVEI, CPVE) and the reform of the 16+ examinations.

B57. NATIONAL CURRICULUM. Task Group on Assessment and Testing (TGAT). *A report of the Task Group*. Chairman: Professor P J Black. Pub'd by the DES and Welsh Office, December 1987.

TGAT was set up (July 1987) to advise the Secretary of State on the practical considerations that should govern all assessment including testing of attainment at age (approx) 7, 11, 14 and 16, within a national curriculum. The supplementary publication *Three supplementary reports* makes recommendations about the various services and arrangements required to support the system proposed. A booklet *A digest for schools* is also available.

B58. NATIONAL MODERN LANGUAGES CONVENTION, 1st, Dublin, 1982.

Innovations in curriculum and examinations at second level: proceedings. Dublin: Trinity College, Centre for Language and Communications Studies, 1982.

Proposals for curriculum development for the teaching of languages in the Republic of Ireland.

B59. OLLER, J W, Jr

Language tests at school: a pragmatic approach. London: Longman, 1979.

Considers the linguistics basis of all tests of achievement, intelligence and aptitude. The focus is on language for purposes of communication. Emphasis on English language teaching.

B60. OLLER, J W, Jr, editor

Issues in language testing research. Rowley, Mass.: Newbury House, (distributed by Harper & Row, UK), 1983.

Updates two earlier volumes: *Language in education: testing the tests* (1978), and *Research in language testing* (1980).

A number of papers in this collection have appeared in other publications.

BIBLIOGRAPHY

B61. OMAGGIO, A C

Proficiency-oriented classroom testing. Washington DC: Center for Applied Linguistics, ERIC Clearinghouse on Language and Linguistics, 1983. (Language in education: theory and practice. Vol 6, No 52).

Part of a project to develop standard, universally acceptable language proficiency guidelines for the language skills and culture across languages.

OSKARSSON, M. See B25 and B26.

B62. PAGE, B, and D HEWETT

Languages step by step: graded objectives in the UK. London: CILT, 1987.

An up-to-date review of developments in graded objectives schemes. Sections cover: Review of progress (syllabus design, assessment, organisation and materials, diversification, statistics); Research (large-scale projects and work carried out by individuals for PhD and other higher degrees); Reports by HMI; Issues (methodology, place of grammar, standardisation); ways forward (pupil profiling, GOML and GCSE, Post 16-TVEI etc); The next step; Bibliography (surveys, periodical articles, reports and projects). Supercedes *Graded objectives in modern languages*, CILT, 1980. See also *GOML News* (Section 4.3).

B63. PORTER, D

A language testing handbook. Oxford: Blackwell. Forthcoming end 1988.

B64. PRICE, Eurwen, and others

Development of assessment materials in the Welsh language (for the primary school). Three reports commissioned by the Welsh Office on an NFER Project. Published 1987, and available from the Welsh Office. 3 Titles: *Reading in Welsh primary schools*; *Welsh reading survey 1984*; *Welsh speaking survey 1985*. A further project, begun November 1987, aims to produce materials to assess Welsh language skills amongst secondary pupils aged 13+. Details from the NFER.

B65. RIVERA, Charles, editor

An ethnographic/sociolinguistic approach to language proficiency assessment. 1983.

BIBLIOGRAPHY

Language proficiency and academic achievement. 1984.

Communicative competence approaches to language proficiency assessment: research and application. 1984.

Placement procedures in bilingual education: education and policy issues. 1984.

The publisher of the above four books is Multilingual Matters, Clevedon, Avon.

The four volumes are comprised of selected papers from the Language Proficiency Assessment Symposium, held March 14-18, 1981, in Warrenton, Virginia, USA. The Symposium was planned as a component of the Assessment of Language Proficiency of Bilingual Persons (ALBPB) project funded by the US National Institute of Education.

B66. STANSFIELD, Charles W, editor

Technology and language testing. Washington, DC: TESOL, 1986.

A collection of papers from the Seventh Annual Language Testing Research Colloquium, held at Educational Testing Service, Princeton, NJ, April 6-9, 1985. Provides an overview of current activity involving the application of computers to the assessment of language proficiency. Section 1 is about computerised adaptive testing and latent trait theory. Section 2 deals with its uses in the development of new measures of speaking, reading and writing.

B67. STEVENSON, D K

Testing, testing. System, Vol 14, No 1, 1986. pp 85-93.

Evaluates five important works on language testing, published 1961-1984. (Finocchiaro, Harrison, Madsen, Oller, Reid).

B68. SMPOSIUM ON THE EVALUATION OF FOREIGN LANGUAGE PROFICIENCY

Proceedings of the symposium; edited by Albert Valdman. Bloomington, Indiana: Indiana University, Committee for Research and Development in Language Construction, 1988.

Symposium held March 4-6, 1987, at Indiana University. Papers mainly on the ACTFL proficiency guidelines.

B69. SUMNER, R

The role of testing in schools. Windsor: NFER-Nelson, 1987.

Against a background of the increasing use of tests this book provides a discussion of the purposes of testing. It examines the advantages and limitations of different types of tests, scales and

their interpretation and the use of test data, and points to future developments. (Publisher).

B70. *TOEFL research reports*. Details: TOEFL Program Office, PO Box 2917, Educational Testing Service, Princeton, N.J., USA.

B71. UNDERHILL, Nic
Testing spoken language: a handbook of oral testing techniques. Cambridge University Press, 1987.

A practical guide to communicative oral testing. Describes the aims and student expectations of testing programmes, the range of test-types and elicitation techniques, marking systems and the evaluation of tests.

4.3 INDEXES, PERIODICALS AND NEWSLETTERS

B72. *APU Newsletter*

Issued at irregular intervals by the DES Assessment of Performance Unit. Distributed free to LEAs, teachers' centres, etc.

B73. *British Education Index*

Published by the University of Leeds, 1986-. Quarterly, with annual cumulation. Begun 1972-. Subject index to articles on education appearing in current issues of over 250 (mainly British) journals. Many articles on language teaching.

B74. *GOML News*

Compiled at CILT for the GOML Co-ordinating Committee, and available from CILT. Annual. Current issue: No 12, Autumn 1987. Provides information about new groups within the graded objectives movement, languages available, group activities, teaching materials and graded tests, teacher-training courses, workshop reports from the yearly course conference, research projects, and bibliographies.

B75. *Language Teaching: The International Abstracting Journal for Language Teachers and Applied Linguists*.

Cambridge University Press. Quarterly. 1968-.

Essential for keeping up-to-date on developments in language teaching and testing.

B76. *Language Testing*

Periodical published by Edward Arnold, London. Vol 1, 1984-. 2

p.a. Publishes articles in the fields of second and foreign language testing, mother tongue testing, and assessment in child language acquisition and language pathology. Includes review articles, reviews of tests as well as of books, research reports, and correspondence.

B77. *Language Testing Update*

A newsletter available from the Institute of English Language Education, University of Lancaster, Lancaster LA1 4YT. 2 p.a. No 1, 1985-. Provides news on new developments in English language testing. Includes reports from examination boards, and from individuals and institutions around the world; descriptions of tests, a calendar of courses and conferences; a bibliography of books and articles; reviews; and abstracts of graduate students' work.

4.4 PERIODICAL ARTICLES (SELECTED TITLES WHICH HAVE NOT APPEARED IN THE ABSTRACTS SECTION)

B78. ALLEN, E D

Communicative competence and levels of proficiency. *Canadian Modern Language Review*, Vol 41, No 6, 1985. pp 991-99.

B79. ALLISON, D, and R Webber

What place for performative tests. *English Language Teaching Journal*, Vol 38, No 3, 1984. pp 199-203.

B80. BOWKER, D

The information gap in placement testing. *English Language Teaching Journal*, Vol 38, No 4, 1984. pp 248-55.

B81. BOYCE, C

The APU Survey - performance in German at 13+. *Treffpunkt*, Vol 17, No 2, June 1985. pp 2-11.

B82. CARROLL, B J

How to develop communicative language tests. *World Language English*, Vol 1, No 1, 1981.

B83. CARROLL, B J

Communicative language tests: tasks, enabling skills, formats and

BIBLIOGRAPHY

measurement criteria. *World Language English*, Vol 2, No 1, 1982. pp 37-39.

B84. CLARK, J L

Classroom assessment in a communicative approach. *British Journal of Language Teaching*, Vol 25, No 1, 1987. pp 9-19.

B85. DAVIES, A

Communicative language testing. *Journal of Applied Linguistics* (Greece), No 1, 1985. pp 22-33.

B86. DAY, E M, and S M Shapson

Assessment of oral communicative skills in early French immersion programmes. *Journal of Multilingual and Multicultural Development*, Vol 8, No 3, 1987. pp 237-60.

B87. DELAMERE, T

Notional-functional syllabi and criterion-referenced tests: the missing links. *System*, Vol 12, No 3, 1984. pp 273-85.

B88. EVANS, C, and others

Testing advanced communicative competence through role-play. *British Journal of Language Teaching*, Vol 25, No 2, 1987. pp 104-8.

B89. GREEN, P S

Reliability in the assessment of written communicative skills. *Finlance*, Vol 3, 1984. pp 1-24.

B90. HAASTRUP, K

Pragmatic and strategic competence in the assessment of oral proficiency. *System*, Vol 14, No 1, 1986. pp 71-9. English as a foreign language in Danish schools.

B91. HAYWARD, T

Passive assessor tests: an alternative to interieurs to assessing communicative performance in spoken English. *World Language English*, Vol 3, No 1, 1983. pp 39-43.

B92. LAESCH, K B, and A van Kleeck

The cloze test as an alternative measure of language proficiency of children considered for exit from bilingual education programs. *Language Learning*, Vol 37, No 2, 1987. pp 171-189.

BIBLIOGRAPHY

B93. *Les Langues Modernes*, Vol 81, No 6, 1987.

Issue on: Le Baccalauréat en Europe.

B94. LEE, B

Performance in Spanish as a first foreign language at age 13. *Vida Hispanica*, Vol 35, No 2, Spring 1985. pp 21-4.

B95. LIGHT, R L, and others

English proficiency and academic performance of international students. *TESOL Quarterly*, Vol 21, No 2, 1987. pp 251-61.

B96. LUKMANI, Y

The communicational testing of reading. *English Language Teaching Journal*, Vol 36, No 4, 1982. pp 217-25.

B97. McCRORY, D P

Orals - the real issue in foreign languages. *Modern Languages*, Vol 68, No 4, 1987. pp 229-35.

B98. MONNANTEUIL, F

L'examen de fin d'études secondaires en France: le Baccalauréat. *Modern Languages*, Vol 68, No 4, 1987. pp 217-21.

B99. MORROW, K

The evaluation tests of communicative performance. *Prospect*, Vol 1, No 2, 1985. pp 32-44.

B100. PALMER, A S

Measures of achievement, communication, incorporation and integration for two classes of formal EFL learners. *RELIC Journal*, Vol 12, No 1, 1981. pp 37-61.

B101. PARTINGTON, J

Towards a science of question-setting in the GCSE. *British Journal of Language Teaching*, Vol 26, No 1, 1988. pp 45-9.

B102. PEACH, L

Graded tests in modern language teaching to adults. *Adult Education*, Vol 58, No 1, 1985. pp 22-4.

B103. PORTER, L P

Using the ACTFL Proficiency Guidelines to achieve goals of

exploratory language courses. *Foreign Language Annals*, Vol 20, No 4, 1987. pp 323-33.

B104. REINWEIN, J

Le test de closure: un outil pour mesurer l'effet de l'illustration sur la compréhension du lecteur. *Revue Québécoise de Linguistique*, Vol 16, No 2, 1987. pp 29-83.

B105. SCHRAFNAGL, J, and D Cameron

Examinations for the 1990s - the Institute of Linguists' syllabus development project. *Modern Languages*, Vol 69, No 1, 1988. pp 43-50.

B106. SHOHAM, M, and others

Reading comprehension tests: general or subject-specific? *System*, Vol 15, No 1, 1987. pp 81-8.

B107. SHOHAMY, E, and others

Introducing a new comprehension test of oral proficiency. *ELT Journal*, Vol 40, No 3, 1986. pp 212-20.

B108. SKEHAN, P

The role of foreign language aptitude in a model of school learning. Comments on Skehan, by A Hughes. *Language Testing*, Vol 3, No 2, 1986. pp 188-223.

B109. SKEHAN, P

Comments on Skehan, by G Sarig. *Language Testing*, Vol 4, No 1, 1987. pp 51-4.

B110. SMITH, M

Assessment of speaking for the Assessment of Performance Unit. *British Journal of Language Teaching*, Vol 22, No 1, 1984. pp 9-14.

B111. SWAIN, M

Large-scale communicative language testing: a case study. *Language Learning and Communication*, Vol 2, No 2, 1983. pp 133-50.

B112. THOMANECK, J K A

Emancipation through communication: the educational aims of CSYS modern language courses. *Modern Languages in Scotland*, No. 23, 1983. pp 96-101.

BIBLIOGRAPHY

B113. TÖNNES-SCHNIER, F, and G Scheibner-Herzig
Measuring communicative effectiveness through dictation. *IRAL*,
Vol 26, No 1, 1988. pp 35-44.

B114. VAN WEEREN, J, and T J J M Theunissen
Testing pronunciation: an application of generalisability theory.
Language Learning, Vol 37, No 1, 1987. pp 109-122/3.

B115. WESCHE, M
Communicative testing in a second language. *Canadian Modern
Language Review*, Vol 37, No 3, 1981. pp 551-71.

5. Tests

5.1 EXAMPLES OF PUBLISHED TESTS

T1. BURT, Maria K, Heidi C Dulay, and Eduardo Hernandez Chavez

Bilingual syntax measure/Medida de sintaxis bilingue (BSM). Manuals 1 and 2. Published by The Psychological Corporation, Harcourt Brace Jovanovich, 1975 (Manual 1) and 1978 (Manual 2). Available through booksellers.

Employs cartoon-type pictures with simple questions in order to identify a student's control over basic syntactic structures in English and Spanish. Suitable for diagnosis or placement.

T2. BRITISH BROADCASTING CORPORATION

Talking of languages - fact pack and aptitude test. 1987. Booklet available from BBC Education.

Linked to the BBC language programmes on Radio 4.

T3. THE BRITISH COUNCIL, and Cambridge Local Examination Syndicate

The English language testing service. Pack consists of: Introduction, new edition, 1987; Sample materials booklet; User handbook; and Wallchart. Further details: English Language Services Department, The British Council, London. (See also Abstracts 27 and 151).

T4. CROUZET, Y, and G Gavelle

Crédif Soltest. Published by Didier/Hatier, Paris, for Centre de Recherche et d'Etude pour la Diffusion du Français, 1979. Available through booksellers.

Test d'évaluation des connaissances en français, langue étrangère, pour l'orientation des adultes (débutants, faux débutants et de niveau perfectionnement). Le test est destiné à mesurer les quatre compétences fondamentales. Consists of: Livret de l'examinateur avec instructions; Feuilles de réponses pour l'élève (in sets).

T5. DUNCAN, D, D Gibbs, N Singh Noor, and H Mohammed Whittaker

Sandwell screening test of expressive language. Published by NFER-Nelson, Windsor, 1987.

TESTS

A picture based test for use with children whose first language is Panjabi. Age range 6-9 years.

T6. INSTITUTE OF LINGUISTS

Bilingual Skills Certificate. 1986.

A nationally recognised qualification in the functional use of two languages, i.e. a community language and English. Testing is available in English with e.g. Arabic, Bengali, Chinese, Greek, Turkish, Vietnamese, etc.

T7. INTERNATIONAL CERTIFICATE CONFERENCE.

Language Certificate Programme.

Examinations are available for Danish, Dutch, English, French, German, Italian, Japanese, Russian, and Spanish. Each syllabus contains an introduction to the examination, a description of the learning objectives, a structure list, a vocabulary list, and a model test. Separate test models and a diagnostic test are available for most languages. The ICC is an international association of adult education organisations which have adopted the language certificate programme developed by the German Adult Education Association which offers examinations throughout Europe at two levels. Further details from the Deutscher Volkshochschulverband, Prüfungszentrale, Eschersheimer Landstr. 61/63. D-6000 FRANKFURT 1. German Federal Republic. (See also Abstract 134).

T8. PHILLIPS, D. and A Miller

Oxford German aptitude test. University of Oxford Department of Educational Studies, 1982.

The test, which uses German as the predictor language, consists of a booklet and recorded material. It is available from CILT at £2.50. Further reading in Oxford Educational Research Group BP Modern Languages Project *Occasional Papers 1 and 2* which are available from the Department.

T9. ROYAL SOCIETY OF ARTS Examinations Board

English as a Second Language Profile Certificate. London: RSA, 1987.

The RSA ESL Profile Scheme is an assessment mechanism which can be applied in a variety of learning environments and which certifies the achievements of each student.

T10. YORK UNIVERSITY. Language Teaching Centre

Language aptitude test.

TESTS

A written test which uses elementary Swedish as the predictor language and tests the candidate's ability to abstract rules from simple examples. Further details: Language Teaching Centre, York University, Heslington, York YO1 5DD.

5.2 COLLECTIONS OF TEST MATERIALS HELD BY THE CILT LIBRARY

Tests

A list of some 25 titles held is available on request. These include the titles listed in Section 5.1.

Graded Objectives in Modern Languages (GOML)

A collection of the unpublished graded syllabuses and tests from all the GOML schemes in operation. Full details about GOML groups is published in *GOML News* (See Section 4.3).

GAMLL Scheme (The joint ILEA/London University School Examinations Board Scheme)

The Graded Assessment in Modern Languages, London, Scheme is in its third year of development and piloting. It aims to set criteria by which pupils can be assessed at any particular time and which are appropriate for their individual rate of progress. The GAMLL scheme will be nationally available in September 1989 and designed to be a direct path to GCSE equivalence. Article in *ILEA Bulletin* Autumn 1987, pp 4-11. Produces the *GAMLL Newsletter*, No 1, 1986-

Teaching Materials Collection

Some language courses have linked achievement tests, which may be graded tests in some cases, e.g. *Eclair*.

Public Examinations

Syllabuses and papers relating to foreign language examinations conducted by British examining bodies. These include GCSE Groups, GCE Boards, Northern Ireland, Scottish and Welsh Boards, Royal Society of Arts, Institute of Linguists, Business & Technician Education Council, London Chamber of Commerce and Industry, Institute of European Education (Lancaster) etc.

International Certificate Conference. Language Certificate Programme

Syllabuses and mock tests are held for Danish, Dutch, English, French, German, Italian, Japanese, Russian, and Spanish. (See Section 5.1).

6. Name index to Abstracts section

This is a name index of *joint authors only* from the Abstracts section. A complete name index for the volume as a whole is not considered necessary as the names within each section are already arranged in alphabetical order.

Bachman, Lyle F	126	Weeks, C Andrew	117
Baker, Rosemary	175	Wolf, B	93
Blok, Henk	137	Woods, Anthony J	103
Blum-Kulka, Shoshana	124		
Burgess, G J A	169		
Cziko, Gary A	67		
Duijkers, Tom	19		
Edwards, Gregg	136		
Frawley, William	104		
Hecht, Karlheinz	76		
Henning, Grant	44,176		
Ilyin, Donna	162		
Jones, Barbara	128		
Kirk, R E	179		
Lynch, Brian	91		
McNally, J	185		
Madsen, Harold S	106		
Nien-Hsuan, Jennifer Lin	43		
Olshtain, Elite	32		
Painchaud, Gisele	107		
Palmer, Adrian S	8,10		
Puntanen, Simo	108		
Ratz, Ulrich	100		
Ramraz, Rachel	15		
Regan, M	88		
Reves, Thea	154		
Savignon, S J	178		
Selinker, Larry	52		
Shikhani, May	86		
Smythe, P C	74		
Underhill, Nick	113		
Vermeer, Anne	171		
Waters, Alan	2		

7. Contents index of abstracts (1-200)

Accelerated learning	146
Acculturation	197
Achievement testing	55, 75, 91, 115, 129, 132, 145, 159, 194
ACTFL	63, 104, 111, 135, 177, 193, 199
Adults	80, 88, 125, 134, 180, 184
Anxiety	199
Aptitude testing	40, 59, 123, 179, 200
Assessment of Performance Unit (DES)	68, 129
Attitude	19, 74, 141
Bilingual education	35, 105, 136, 143, 167, 181
Bilingualism	130
C-tests	100, 101
Cloze procedures	1, 9, 11, 15, 22, 23, 29, 30, 51, 62, 64, 65, 78, 81, 92, 100, 101, 112, 144
Comics	197
Common standard	58
Communicative competence	10, 26, 27, 28, 56, 62, 68, 77, 80, 82, 94, 95, 109, 116, 119, 120, 124, 129, 136, 140, 143, 150, 154, 167, 177, 180, 183
Comparison of methods	84, 113, 153

CONTENTS INDEX OF ABSTRACTS

Competitions	173
Comprehension, listening	4, 49, 50, 85, 191
Comprehension, reading	12, 13, 14, 15, 16, 53, 78, 96, 128, 153, 185, 186, 195
Computer methods	31, 34, 42, 47, 53, 106
Continuous assessment	70, 146
Cross-cultural testing	78, 124, 181, 185, 200
Diagnostic testing	20, 47, 114, 121, 123, 178
Dialogue	72
Dictation	5, 42, 67, 143, 164
Dictionaries	13, 14
Difficulty	12, 44, 190
Disclosure	79
Dutch	102, 171, 181
English (EFL)	1, 4, 12, 13, 14, 15, 20, 21, 22, 30, 37, 42, 43, 44, 45, 46, 49, 50, 56, 61, 67, 79, 81, 83, 84, 85, 86, 90, 91, 92, 97, 102, 108, 113, 115, 119, 121, 123, 127, 128, 131, 134, 144, 145, 150, 151, 163, 176, 178, 180, 184, 190, 194, 196, 200
English for special purposes	2, 27, 60, 114, 157, 158, 168, 182, 191, 195
English - (ESL/UK)	45, 48, 56, 155, 170, 185
English mother tongue	48, 57

CONTENTS INDEX OF ABSTRACTS

Elicited imitation	80
Error analysis	105
Evaluation of tests	45, 46, 67, 93, 113, 119, 126, 129
Examinations	21, 25, 45, 48, 55, 60, 69, 93, 99, 134, 139, 155, 168, 170, 174, 187, 189, 198
French	7, 18, 34, 38, 62, 77, 94, 105, 116, 120, 141, 143, 146, 167, 179, 183
German	16, 53, 63, 80, 112, 117, 135, 169
Grammar	20, 190
Graded tests	24, 37
Group testing	18, 26, 108, 138
History	164
Hebrew	152
Holiday courses	113
Immigrants	39, 48, 88, 125, 180
Individualised instruction	88, 147
Intelligibility	121
Interlanguage	52
Interview technique	8, 58, 60, 63, 98, 111, 178
Italian	69
Language for special purposes	118, 135, 136, 146
Language laboratory	72, 169

CONTENTS INDEX OF ABSTRACTS

Latent trait methods	86, 87, 106
Listening comprehension, <i>see</i> Comprehension, listening	
Low proficiency learners	24, 85
Macrosimulation	88
Multiple-choice	12, 15, 34, 114, 185
Needs analyses	88
Noise masking	5
Objective methods	6, 54
Oral fluency	166
Performance	35, 68, 71, 97
Personality	145
Primary schools	56, 132, 185
Placement testing	27, 107, 114, 117, 144, 176
Portuguese	51
Pragmatic test	94, 124
Proficiency testing	7, 17, 18, 43, 46, 58, 61, 81, 83, 84, 85, 98, 104, 122, 131, 135, 142, 163, 171, 172, 177, 196, 199
Profiling	57
Psychological tests	39
Radio recordings	50
Reading comprehension <i>see</i> Comprehension, reading	

CONTENTS INDEX OF ABSTRACTS

Review of literature	4, 31, 142
Reliability	76, 103
Secondary schools	19, 20, 21, 24, 25, 38, 55, 57, 66, 76, 97, 93, 102, 116, 120, 123, 129, 141, 167, 173, 178, 183, 187, 189, 192, 198
Self-assessment	35, 44, 89, 107, 125, 133, 137, 150
Sex differences	108
Social correlates	61, 66, 71, 162
Sociocultural competence	32
Spanish	159
Speech production	6, 18, 21, 22, 57, 60, 63, 69, 84, 93, 98, 104, 108, 135, 138, 139, 152, 163, 171, 174, 194, 196, 198, 199
Speech variety	66
Substitution tests	36
Surveys of tests	55, 97, 154, 155
Syntax	73, 184
Teacher training	2, 99, 116, 139, 165
Teaching materials	38, 120
Test design	3, 10, 11, 21, 27, 49, 74, 75, 95, 98, 109, 110, 111, 115, 119, 136, 138, 151, 163
Theory and principles	3, 17, 25, 26, 28, 31, 33, 41, 52, 64, 73, 82, 100, 101, 103, 126,

CONTENTS INDEX OF ABSTRACTS

	137, 140, 148, 154, 157, 158, 160, 161, 165, 172, 175, 192
TOEFL	79, 131, 191
Translation	99, 102
Turkish	171
Unidimensionality	87
United Kingdom	24, 25, 38, 48, 68, 69, 76, 116, 119, 129, 132, 141, 169, 170, 174, 183, 187, 189, 198
Welsh	132
Writing	76, 90, 110, 122, 127, 137, 149, 188

This supplement consists of abstracts from the journals *Language Teaching and Linguistics: Abstracts* and *Language Teaching* (1981-1987). Also included are an annotated listing of survey articles, books, etc published mainly during the same period, and a detailed subject index and name index.

During this time interest in testing has focused mainly on the problems involved in specifying communicative proficiency and in reshaping of the public examination system at 16 + .

CILT